De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen

Uiterwijk, Jan Hendrik

Publication date:
1994

Document Version
Publisher's PDF, also known as Version of record

Eindtoets
Basisonderwijs

Henny Uiterwijk

De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen

Cito

De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen

De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Brabant,
op gezag van de rector magnificus,
prof. dr. L.F.W. de Klerk,
in het openbaar te verdedigen
ten overstaan van een
door het college van dekanen
aangewezen commissie
in de aula van de Universiteit
op vrijdag 20 mei 1994 te 16.15 uur

door

Jan Hendrik Uiterwijk

geboren te Arnhem

Promotores: Prof. dr. A.J.A.G. Extra
Prof. dr. L.F.W. de Klerk
Co-promotor: Dr. A. Vallen

Omslagontwerp en grafische vormgeving: Hélène de Wit

Voorwoord

Om na te gaan of de Eindtoets Basisonderwijs voor allochtone leerlingen even goed bruikbaar is als voor autochtone, hebben medewerkers van het Werkverband Taal en Minderheden van de Letterenfaculteit van de Katholieke Universiteit Brabant (KUB) en medewerkers van het project Eindtoets Basisonderwijs van het Instituut voor Toetsontwikkeling (Cito) samen een onderzoeksproject uitgevoerd. In dit project zijn de toetsscores van allochtone en autochtone leerlingen op (onderdelen van) de Eindtoets Basisonderwijs berekend en is vastgesteld hoe van deze leerlingen de toelating tot en de doorstroming in het voortgezet onderwijs verloopt. Verder is onderzocht of enerzijds de Eindtoets Basisonderwijs als geheel en anderzijds de afzonderlijke toetsitems ook voor allochtone leerlingen aan hun doel beantwoorden. In het eerste geval gaat het om de vraag of er al dan niet sprake is van toetsbias, in het laatste geval gaat het om itembias.
Toetsbias wordt in dit verband opgevat als onderzoek naar de vraag hoe hoog de voorspellende waarde van de Eindtoets Basisonderwijs is voor allochtone en autochtone leerlingen in vergelijking met de voorspellende waarde van het schoolkeuze-advies van de basisschool. In het onderzoek naar itembias zijn twee complementaire fasen onderscheiden. In de eerste fase zijn met statistische procedures items opgespoord waarbij sprake is van itembias. In de tweede fase is een poging ondernomen om te onderzoeken wat bij een bepaald item de oorzaak van itembias zou kunnen zijn. Bij het achterhalen van die mogelijke oorzaken van itembias zijn drie groepen personen betrokken geweest: de projectmedewerkers (van KUB en Cito), niet bij het onderzoeksproject betrokken experts en leerlingen uit groep acht van het basisonderwijs.
Na de start van het project bleek al gauw dat het onderzoek naar itembias in meerdere opzichten een ontdekkingsreis zou worden. Zo werd bijvoorbeeld spoedig duidelijk dat in de Verenigde Staten weliswaar veel aandacht is besteed aan statistische procedures voor het opsporen van itembias, maar tevens bleek daarbij dat vergelijkbare procedures niet tot dezelfde resultaten leiden. Op de vraag bij hoeveel items van een bepaalde toets sprake is van itembias, zijn dan ook verschillende antwoorden mogelijk. Verder bleek dat met het zoeken naar oorzaken van itembias, niet alleen in Nederland maar ook in andere landen, bijzonder weinig ervaring is opgedaan. Goed gefundeerde taalkundig-inhoudelijke verklaringen inzake itembias voor allochtone leerlingen ontbreken geheel. Omdat een theoretisch kader betreffende bronnen van itembias voor allochtone leerlingen vooralsnog niet voorhanden is, hebben de conclusies, die op basis van het onderhavige onderzoek in dit verband worden getrokken, een voorlopig karakter.

Bij de uitvoering van het onderzoeksproject zijn in de verschillende jaren diverse personen betrokken geweest. Zonder hun inzet en stimulerende invloed zou dit project wellicht nooit tot een goed einde zijn gebracht. Een aantal mensen wil ik hier in het bijzonder bedanken.
In de eerste plaats noem ik dr. Ton Vallen (KUB) die van begin tot eind op uiterst constructieve wijze het project aan Tilburgse zijde heeft geleid. In de beginfase van het project heeft ook dr. Anne Kerkhoff (KUB) bij de vragenlijst-constructie en bij de eerste zoekpogingen naar mogelijke oorzaken van itembias

veel waardevolle suggesties gedaan. Bij de vragenlijstconstructie is ook dankbaar gebruik gemaakt van de expertise van drs. Fons Moelands (Cito). Drs. Marianne de Jong en drs. Marja Coenen hebben als Assistenten In Opleiding van de Letterenfaculteit van de KUB een belangrijke bijdrage geleverd aan het project. Dat geldt vooral ten aanzien van het achterhalen van mogelijke bronnen van itembias. Door het aanvaarden van een werkkring elders hebben beiden helaas vroegtijdig hun werkzaamheden beëindigd. Drs. Marijke van de Waal (KUB) heeft als student(assistente) het onderzoek uitgevoerd naar de oordelen van experts over bronnen van itembias, waarover ze in haar doctoraalscriptie verslag heeft gedaan. Dr. Ron Engelen (Cito) heeft het project terzijde gestaan met methodologische adviezen, ook op het relatief nieuwe terrein van itembias.

Deze dissertatie vormt het eindverslag van het genoemde samenwerkingsproject van het Werkverband Taal en Minderheden en het Cito. Bij de totstandkoming van de dissertatie hebben een aantal mensen, ondanks hun vele andere werkzaamheden, bijzonder waardevolle ondersteuning verleend. Ik ben hen zeer veel dank verschuldigd.
De stimulerende invloed die bij het schrijven van deze dissertatie van de co-promotor dr. Ton Vallen is uitgegaan, kan moeilijk overschat worden. Hij heeft steeds op vriendschappelijke wijze duidelijk gemaakt wat er nog kon en nog moest gebeuren. Van zijn grote kennis en inzicht op het terrein van de linguïstiek en de taalvaardigheid van allochtone leerlingen heb ik veel geleerd.
De gesprekken met de beide promotores, prof. dr. Guus Extra en prof. dr. Len de Klerk, hebben er toe geleid dat het aantal blinde vlekken op mijn netvlies is verminderd. Het was een genoegen om met zulke breed georiënteerde mensen over het manuscript te kunnen discussiëren.
Dr. Johan Wijnstra (Cito) heeft het manuscript eveneens van kritisch commentaar voorzien. Ik heb niet alleen in verband met mijn dissertatie een beroep mogen doen op zijn grote kennis en inzicht als onderwijskundig onderzoeker, maar ik heb het voorrecht in hem al bijna 15 jaar een uitstekende collega te hebben die steeds bereid is te luisteren en te adviseren.
Zowel bij de uitvoering van het project als het schrijven van de dissertatie heb ik veel geleerd op methodologisch terrein. Dr. Ron Engelen (Cito) bleef steeds geduldig uitleggen welke procedures gevolgd moesten worden en waarom dat moest. Annelies van Exter (Cito) heeft samen met de Grafische Dienst van het Cito ervoor gezorgd dat een diskette met verschillende soorten bestanden is omgewerkt tot dit fraaie boek.

Het schrijven van een dissertatie wordt gemakkelijker gemaakt door een stimulerende omgeving. Mensen die niet bij het schrijven zelf betrokken zijn, maar die laten merken dat ze dit soort werk waarderen en die van tijd tot tijd informeren naar de voortgang. In dit verband wil ik twee groepen mensen in het bijzonder bedanken: mijn collega's en mijn gezinsleden.
De medewerkers van het Cito en met name die van de sector Basis- en Speciaal Onderwijs vinden het vanzelfsprekend dat je van je werk schriftelijk verantwoording aflegt en ze vinden het niet ongebruikelijk om dat te doen in de vorm van een dissertatie. Deze opvatting brengt mensen ertoe hoge eisen te stellen aan hun werk.

Mijn vrouw Harmke en onze zonen Rik en Koen hebben mij de afgelopen tijd enerzijds de nodige ruimte gegeven en anderzijds bleven ze mij bij het gezinsleven betrekken. Zodoende hebben ze ertoe bijgedragen dat mijn studeerkamer geen isoleercel is geworden.

Zelhem, februari 1994

Inhoud

1 De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen

1.1 Inleiding

Het is niet ongebruikelijk om voor de beschrijving van de schoolprestaties van allochtone en autochtone leerlingen toetsresultaten te gebruiken, zeker als het erom gaat de verschillen tussen beide groepen in de beheersing van een aantal aspecten van de Nederlandse taal tot uitdrukking te brengen. Ook voor de communicatie met personen buiten het onderwijsveld worden toetsresultaten gebruikt. Zo hanteren Tesser & Vierke (1990) als operationalisering van schoolprestaties onder andere scores op taaltoetsen, wanneer zij op verzoek van het Ministerie van Binnenlandse Zaken rapporteren over de schoolprestaties van allochtone en autochtone leerlingen in het basisonderwijs.

Tegen de achtergrond van het bovenstaande wekt het enige verbazing dat er tot nu toe in Nederland nauwelijks onderzoek is gedaan naar de vraag of veelgebruikte toetsen wel een geschikt middel zijn om de vaardigheid van zowel allochtone als autochtone leerlingen op het terrein van bepaalde onderwijsdoelstellingen te meten. Vooral wanneer de gemiddelde toetsscores van onderscheiden groepen, zoals allochtone en autochtone leerlingen, aanzienlijk verschillen, kan de onderzoeker zich immers afvragen of die verschillen toe te schrijven zijn aan verschillen in de te meten vaardigheden of dat ze een artefact zijn van de gehanteerde meetprocedure.
Sommige onderzoekers verwachten dat toetsen een onderschatting geven van het prestatieniveau van allochtone leerlingen. Vallen & Kerkhoff (1985) zijn bijvoorbeeld van mening dat van alle leerlingen de resultaten op de gebruikelijke toetsen met de nodige reserves bekeken moeten worden. Volgens hen gelden die reserves ten aanzien van allochtone kinderen in nog sterkere mate. De linguïstische en culturele achtergronden van deze leerlingen spelen volgens hen een belangrijke rol bij het maken van toetsen in het immigratieland.
Ook vanuit de onderwijspraktijk wordt de bruikbaarheid van toetsen voor leerlingen uit etnische minderheidsgroepen zo nu en dan betwijfeld. De Turkse Leerkrachten Vereniging in Gelderland stelt bijvoorbeeld dat het afnemen van toetsen, waaronder de Cito-toetsen in groep acht van het basisonderwijs, bij allochtone leerlingen in de praktijk veel problemen oplevert. De vraagstelling van de Cito-toets is volgens hen vaak cultureel bepaald, waardoor van allochtone leerlingen meer wordt gevraagd dan van autochtone leerlingen (Ersoy, 1991).
Er worden door onderzoekers pogingen in het werk gesteld om anderen ervan te overtuigen dat de door hen gehanteerde toetsen voor allochtone leerlingen bruikbaar zijn. Zo stelt Driessen (1990: 74) dat het toetsen van allochtone leerlingen in een taal die voor een groot deel van hen niet de moedertaal is, mogelijk een probleem vormt. Bij de constructie van de in zijn onderzoek gebruikte toetsen is daar dan ook, volgens hem, speciale aandacht aan besteed. Hij deelt evenwel niet mee welke maatregelen er genomen zijn om de te meten vaardigheden bij allochtone leerlingen adequaat te meten. Driessen (1990: 206) concludeert echter ook dat het erop lijkt dat de methodiek om te bepalen of een

toets bruikbaar is voor allochtone leerlingen nog in de kinderschoenen staat.

De twijfel aan de bruikbaarheid van meetinstrumenten voor leerlingen uit etnische minderheidsgroepen geldt ook voor intelligentietests. Extra & Verhoeven (1985) zijn bijvoorbeeld van mening dat een intelligentietest die bedoeld is voor monolinguale leerlingen, niet zonder meer te gebruiken is om de intelligentie te meten van allochtone kinderen. Zowel de Nederlandstalige instructie bij de nonverbale en verbale taken als de verbale taken zelf houden geen rekening met de meertalige achtergrond van allochtone kinderen. Er moet, volgens Extra & Verhoeven (1985), rekening mee gehouden worden dat met de tot nu toe gebruikte intelligentietests niet nagegaan wordt hoe intelligent allochtone leerlingen zijn, maar in welke mate ze het Nederlands als tweede taal beheersen. Van de Vijver (1991: 66) stelt dat een test bij onderscheiden culturele groepen hetzelfde psychologische construct moet meten. Na een beschrijving van de literatuur over 'culture-fair' tests komt hij (1991: 65) tot de conclusie dat ook als vooraf eisen zijn geformuleerd om de bruikbaarheid van een test bij onderscheiden culturele groepen te maximaliseren, de implementatie niet garandeert dat meetartefacten, bijvoorbeeld veroorzaakt door differentiële vertrouwdheid met het stimulusmateriaal, geëlimineerd zijn. Hofstee, voorzitter van een testscreeningscommissie die twintig van de in Nederland meest gebruikte psychologische tests op 'cultural bias' en op cultuurgebonden en racistische items doorlichtte, komt tot de conclusie dat de schijnbaar eenvoudige vraag of een test allochtone leerlingen benadeelt, in feite een gecompliceerde kwestie is en dat het empirisch onderzoek in deze aan hoge eisen moet voldoen (Hofstee, 1990). Deze testscreeningscommissie spoort tot verhoogde onderzoeksinspanning op dit terrein aan en deponeert de bewijslast voor testfairness, in de zin van afwezigheid van testbias, bij de testontwikkelaar en testgebruiker.

In de zomer van 1985 hebben medewerkers van het Werkverband Taal en Minderheden van de Letterenfaculteit van de Katholieke Universiteit Brabant (KUB) contact gezocht met medewerkers van het project Eindtoets Basisonderwijs van het Instituut voor Toetsontwikkeling (Cito) om de mogelijkheden te verkennen samen een onderzoeksproject op te zetten.
Het ging toen vooral om de volgende onderzoeksvragen:
- Hoe ontwikkelen de scores op de Eindtoets Basisonderwijs van allochtone en autochtone leerlingen zich in de komende jaren?
- Met welke items en toetsonderdelen hebben allochtone leerlingen specifieke problemen?
- Welke mogelijkheden zijn er om de Eindtoets zo aan te passen, dat eventuele 'biases' voor allochtone leerlingen in verband met hun talige en culturele achtergrond opgeheven worden?

De KUB en het Cito besloten samen een onderzoek te gaan uitvoeren om antwoorden te vinden op bovenstaande en een aantal aanvullende onderzoeksvragen. Er werd afgesproken om, nadat in 1986 een vooronderzoek zou zijn gehouden, achtergrondgegevens te verzamelen van de leerlingen die in 1987 en 1989 aan de Eindtoets Basisonderwijs zouden deelnemen.

De keuze om de Eindtoets Basisonderwijs te laten fungeren als object van onderzoek is vooral ingegeven door het feit dat elk jaar een groot aantal leerlingen aan deze toets deelneemt. In de periode 1980 – 1990 ligt het aantal deelnemers aan de toets tussen de 75 000 en 100 000. In 1987 namen 3801 scholen met samen 80 685 leerlingen aan de toets deel; in 1989 waren dit 4652 scholen met 92 448 leerlingen. Dat is ongeveer 45%, respectievelijk 55% van het totaal aantal leerlingen in groep acht van het basisonderwijs. Sinds 1992 ligt het aantal Eindtoetsdeelnemers zelfs boven de 100 000.
De Eindtoets Basisonderwijs, waarvan elk jaar een nieuwe versie verschijnt, heeft twee functies. Enerzijds verschaft de toets informatie over individuele leerlingen in verband met de overgang naar het voortgezet onderwijs, anderzijds levert de toets informatie voor de evaluatie van het onderwijsprogramma van de basisschool. In het onderhavige onderzoek staat de eerste functie centraal.
De toets bestaat uit 180 opgaven die evenredig verdeeld zijn over de onderdelen Taal, Rekenen en Informatieverwerking. De inhoud van de toets wordt verantwoord in het zo geheten Doelenboek, de inhoudsverantwoording van de Eindtoets Basisonderwijs (Cito, 1986a). Op leerlingniveau wordt gerapporteerd over het totaal en op het niveau van de toetsonderdelen Taal, Rekenen, Informatieverwerking.
Om de scores van een toets die moet functioneren voor de keuze van een school voor voortgezet onderwijs, te kunnen interpreteren, moet de relatie gelegd kunnen worden tussen de scores en de verschillende typen voortgezet onderwijs. Bij de Eindtoets Basisonderwijs gebeurt dit door toelatings- en doorstroomgegevens te verstrekken van leerlingen die in een voorgaand jaar aan de toets deelnamen. Aan de hand van de behaalde totaalscore, die door de zo geheten equivaleringsprocedure (zie 3.1.2) van jaar tot jaar vergelijkbaar is, wordt de positie geschat die de leerling in de verschillende typen voortgezet onderwijs zal innemen als de leerling naar dat type zou gaan. Deze schatting is gebaseerd op onderzoek naar de scoreverdeling in de diverse typen voortgezet onderwijs (Cito, 1988b; Engelen & Uiterwijk, 1990; Cito, 1990; Uiterwijk & Engelen, 1992). Figuur 1.1 geeft een voorbeeld van het leerlingrapport van de Eindtoets Basisonderwijs 1987 en 1989.

Figuur 1.1 Leerlingrapport Eindtoets Basisonderwijs

Cito Centraal Instituut voor Toetsontwikkeling

Resultaten	Taal	Rekenen	Informatieverwerking	Totaal
Aantal opgaven	60	60	60	180
Aantal goed	43	37	50	130
Percentielscore	36	33	69	43

Standaardscore 535

Onderstaand figuur geeft u een indruk of de leerling met deze standaardscore waarschijnlijk tot de betere, zwakkere dan wel tot de middenmoot behoort in een bepaald schooltype. De poppetjes achter de schooltypes geven alle leerlingen weer die naar dat schooltype gaan. Het gemerkte poppetje geeft de plaats aan die de leerling inneemt tussen zijn medeleerlingen in dat schooltype.

	% met eenzelfde of lagere score	0% 20% 40% 60% 80% 100%	% met hogere score
LBO	83		17
MAVO	53		47
MAVO/HAVO en MAVO/HAVO/VWO	29		71
HAVO,VWO en HAVO/VWO	5		95

Nader onderzoek naar de bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen wordt vooral ingegeven door de wens meer duidelijkheid te verwerven over het meten van vaardigheden bij een doelgroep waarvan de sociaal-culturele en linguïstische achtergrond over het algemeen sterk verschilt van die van autochtone leerlingen en waarvan bovendien bekend is dat ze bij meting van verschillende vaardigheden lagere scores behalen. Empirisch onderzoek moet duidelijk maken of de scores op de Eindtoets Basisonderwijs een over- of onderschatting of een juiste weergave geven van de vaardigheid van allochtone leerlingen in de gemeten domeinen.

Het samenwerkingsproject KUB – Cito richt zich op drie onderdelen.
Ten eerste heeft het onderzoek betrekking op het beschrijven van trends in de schoolresultaten van allochtone en autochtone leerlingen. Met schoolresultaten worden hier de toetsscores van deze leerlingen op de (onderdelen van de) Eindtoets Basisonderwijs bedoeld en de gegevens over toelating tot en doorstroming in het voortgezet onderwijs. In de tweede plaats gaat deze studie over onderzoek naar toetsbias. Toetsbias wordt hier opgevat als onderzoek naar de vraag hoe hoog de voorspellende waarde van de Eindtoets Basisonderwijs is voor allochtone en autochtone leerlingen in vergelijking met die van het advies van de basisschool. Het derde onderdeel gaat over het onderzoek naar itembias. In het onderhavige onderzoek naar itembias worden twee complementaire fasen onderscheiden. In de eerste fase worden met statistische procedures items opgespoord waarbij sprake is van itembias. In de tweede fase wordt ingegaan op de vraag wat bij een bepaald item de oorzaak van itembias zou kunnen zijn. Bij het opsporen van mogelijke oorzaken van itembias werden drie groepen personen betrokken, respectievelijk de projectmedewerkers van KUB en Cito, niet bij het project betrokken experts en leerlingen uit groep acht van het basisonderwijs. In 1.3.3 wordt de opzet van het onderzoek naar itembias nader beschreven. Uit het onderzoek naar toets- en itembias moet ook blijken met welke aanpassingen de bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen eventueel vergroot kan worden.
Deze dissertatie vormt het eindverslag van dit onderzoek. In de volgende paragraaf (1.2) wordt ingegaan op het begrippenpaar toets- en itembias, terwijl dit hoofdstuk wordt afgesloten met een overzicht van de centrale onderzoeksvragen (1.3). In hoofdstuk twee worden mogelijke oorzaken van bias voor leerlingen uit etnische minderheidsgroepen aan de orde gesteld. In hoofdstuk drie staat de beschrijving en verantwoording van de gebruikte onderzoeksinstrumenten centraal. In hoofdstuk vier komen de trends in de toetsscores van de onderscheiden etnische groepen op de (onderdelen van de) Eindtoets Basisonderwijs 1987 en 1989 aan de orde. Bovendien worden de trends in de toelatings- en doorstroomgegevens van deze leerlingen in het voortgezet onderwijs gegeven. In hoofdstuk vijf wordt verslag gedaan van het onderzoek naar de predictieve validiteit van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen in vergelijking met die van het advies van de basisschool. Hoofdstuk zes heeft betrekking op de procedures die gevolgd zijn om items op te sporen waarbij sprake is van bias. De resultaten van de verschillende analyses worden besproken en vergeleken. In hoofdstuk zeven wordt aangegeven welke verklaringen er mogelijk te geven zijn voor itembias. In het laatste hoofdstuk wordt een samenvatting gegeven en worden de conclusies van het gehele onderzoek gepresenteerd.

1.2 Toets- en itembias

Bij de beoordeling van de kwaliteit van toetsen is de vraag naar de validiteit een centraal punt. In validiteitsonderzoek wordt nagegaan in welke mate een toets aan zijn doel beantwoordt (Drenth, 1973). Dergelijk onderzoek moet resulteren in een oordeel over de mate waarin er empirische evidentie bestaat voor de bewering dat scores bepaalde conclusies en acties toelaten (vgl. Jensen, 1980; Messick, 1986; 1987). De validiteit heeft dus betrekking op het gebruik van de toetsresultaten. Een toets kan voor een bepaald doel zeer valide zijn, maar niet voor een ander doel. De items van een toets lokken bij de toetsdeelnemer bepaalde responsen uit en aan de hand van deze responsen wordt de status van de toetsdeelnemer ten opzichte van een bepaald construct of criterium vastgesteld. De toetsontwikkelaar heeft tot taak de relatie tussen de scores op een verzameling items en het construct of criterium te verantwoorden.

Validiteitsonderzoek kan verschillende vormen aannemen. De 'American Educational Research Association' (AERA), de 'American Psychological Association' (APA) en de 'National Council on Measurement in Education' (NCME) hebben in een gezamenlijke publicatie (1985) een indeling naar drie soorten validiteit gegeven.

- Inhoudsvaliditeit wordt geëvalueerd door vast te stellen hoe goed de inhoud van een toets het domein van situaties, kennisinhouden of vaardigheden representeert waarover conclusies getrokken moeten worden.

- Criteriumvaliditeit wordt geëvalueerd door de scores te vergelijken met een externe variabele, die verondersteld wordt een directe meting te zijn van het gedrag in kwestie. Er worden twee soorten criteriumvaliditeit onderscheiden:
  - Predictieve validiteit die de mate aangeeft waarin een score iemands toekomstige niveau op een criterium kan voorspellen.
  - Gelijktijdige (concurrent) validiteit die de mate aangeeft waarin een score iemands huidige niveau op een criterium kan schatten.

- Constructvaliditeit wordt geëvalueerd door te onderzoeken welke psychologische kwaliteiten een toets meet. 'Construct' is dan een gepostuleerde vaardigheid waarvan verondersteld wordt dat deze gereflecteerd wordt in de toetsprestatie.

Messick (1987) benadrukt dat deze drie soorten validiteit niet gezien moeten worden als alternatieven, maar als aspecten van validiteitsonderzoek. Hij wijst op de overeenkomst tussen criterium- en constructvaliditeit. In onderzoek naar de criteriumvaliditeit van een meetinstrument is het immers essentieel om te bepalen in hoeverre de externe variabele (het criterium) hetzelfde meet als het meetinstrument in kwestie beoogt te meten. Messick benadrukt dat het bij criteriumvaliditeit niet enkel en alleen gaat om de correlatie tussen toets en criterium. Het is van belang om te verklaren waarom er een bepaald verband bestaat tussen het criterium en de toets (vgl. ook Cronbach, 1972; Drenth, 1972). Hiervoor moeten het criterium en het meetinstrument in kwestie geanalyseerd worden en moet onderzocht worden wat de invloed van relevante variabelen op het meetinstrument en op het criterium is. Dit betekent

onderzoek naar de constructvaliditeit van toets en van criterium.
Volgens Jensen (1980) constateerden Binet en Simon reeds dat hun intelligentietest, ontwikkeld voor Parijse arbeiderskinderen, afgenomen bij kinderen met een hogere sociaal-economische status aanzienlijk hogere gemiddelde testscores opleverde. Vertegenwoordigden de scores van de lagere en hogere sociale milieus inderdaad verschillende intelligentieniveaus of waren de verschillen een artefact van de test? Binet heeft deze onderzoeksvraag nooit formeel onderzocht (Jensen, 1980), maar niet direct verklaarbare verschillen tussen relevante geledingen in de populatie zijn vaak aanleiding om te onderzoeken of een test of toets ook voor onderscheiden subpopulaties aan zijn doel beantwoordt.

Zo wordt er in Nederland bijvoorbeeld nagegaan of de items van het Centraal Schriftelijk Eindexamen moderne vreemde talen van het LBO, MAVO, HAVO en VWO voor jongens en meisjes op dezelfde wijze functioneren. Bij deze examens worden items opgespoord die de leden van de ene sekse significant beter maken dan de leden van de andere met een vergelijkbaar gemiddeld prestatieniveau (Bügel & Robben-Willems, 1989; Bügel, 1991; Bügel & Glas, 1991). In de Verenigde Staten besteedt men op soortgelijke wijze veel aandacht aan de validiteit van toetsen voor leerlingen uit etnische minderheidsgroepen in vergelijking met die voor de blanke meerderheidsgroep (Berk, 1982; Holland & Wainer, 1993). Tatsuoka e.a. (1988) gingen na of een toets ook aan zijn doel beantwoordt voor leerlingen van een vergelijkbaar prestatieniveau die bij bepaalde cognitieve taken verschillende 'problem-solving'-strategieën hanteren.

In onderzoek naar de validiteit van toetsen voor subgroepen wordt het begrip 'bias' gehanteerd. In het algemeen verwijst bias naar de systematische over- of onderschatting van een parameter als functie van het lidmaatschap van een onderscheiden subgroep (vgl. Jensen, 1980; Reynolds, 1982). Biasonderzoek kan betrekking hebben op de toets als geheel en op de afzonderlijke toetsopgaven. Bij een toets die gebruikt wordt om iemands niveau op een extern criterium te schatten, kan onderzocht worden of de criteriumvaliditeit voor de onderscheiden subgroepen even hoog is. Biasonderzoek heeft betrekking op de constructvaliditeit, wanneer onderzocht wordt of de afzonderlijke toetsitems voor de onderscheiden subgroepen het construct op dezelfde wijze representeren.

Elk onderzoek naar de bruikbaarheid van toetsen voor relevante geledingen in de populatie is nog geen toets- of itembiasonderzoek. Hofstee (1990) maakt onderscheid tussen onderzoek naar bias en onderzoek waarin beoordeeld wordt of er in een toets of test etnocentrische of racistische inhouden voorkomen. Volgens Hofstee is de vraag of een tekst of een afbeelding door de beugel kan een kwestie van oordeelsvorming. De vraag of een leerling door bepaalde plaatjes of bewoordingen benadeeld wordt, is daarbij niet aan de orde. Alleen empirisch onderzoek kan uitsluitsel geven of scores van leerlingen door bepaalde inhouden beïnvloed worden: "Een test kan etnocentrische inhoud vertonen of niet, en los daarvan allochtonen benadelen of niet. De beide criteria zijn onafhankelijk van elkaar" (Hofstee, 1990: 292). Ekstrom, Lockheed & Donlon (1979) daarentegen spreken over onderzoek naar 'bias' wanneer de inhoud van een test geanalyseerd wordt. Zo stellen zij bijvoorbeeld dat er bij

een test sprake is van 'bias', wanneer er in de testinhoud vaker mannelijke zelfstandige naamwoorden voorkomen dan vrouwelijke.

In deze dissertatie wordt aangesloten bij de opvatting van Hofstee (1990). Met 'onderzoek naar bias' wordt verwezen naar empirisch onderzoek waarbij nagegaan wordt of het item of de toets het te meten construct, respectievelijk criterium voor onderscheiden subgroepen vergelijkbaar representeren. Indien de items bij bepaalde subgroepen iets anders meten, kan dit de scores van die groepen beïnvloeden. Voor het beoordelen van de inhoud van toetsen of testen op zich wordt hier de term 'inhoudsanalyse' gehanteerd.

'Bias' is niet hetzelfde als 'moeilijkheid'. Regelmatig blijkt dat de resultaten van verschillende bevolkingsgroepen op toetsen verschillen. Op zich is dit geen argument om aan de kwaliteit van de toets te twijfelen. We moeten er altijd rekening mee houden dat ene bevolkingsgroep gemiddeld vaardiger is in het te meten construct dan de andere. Als bijvoorbeeld taalitems voor bepaalde leerlingen moeilijker zijn dan voor andere, wordt meestal voldaan aan de functie van die items of de taaltoets als geheel: het discrimineren tussen meer en minder taalvaardige leerlingen met betrekking tot de taal die getoetst wordt.

Er wordt afbreuk gedaan aan de constructvaliditeit van het meetinstrument wanneer voor het juist beantwoorden van de items nog andere vaardigheden nodig zijn dan de vaardigheid die de items beogen te meten. Wanneer de benodigde additionele vaardigheden niet bij alle onderscheiden subgroepen in vergelijkbare mate aanwezig zijn, spreken we van bias. Dat kan bijvoorbeeld het geval zijn wanneer het niet tot het te meten construct behorende taalgebruik in een rekenopgave voor een bepaalde groep leerlingen dermate ingewikkeld is, dat ze ten gevolge daarvan niet aan het uitvoeren van de beoogde rekenoperatie toekomen of daaraan onvoldoende aandacht kunnen besteden. De vaardigheid die de toetsitems beogen te meten, spelen bij onderzoek naar bias een cruciale rol.

Kok (1988) hanteert als equivalent voor 'bias' het Nederlandse begrip 'partijdigheid'. In navolging van hem worden in dit proefschrift 'bias' en 'partijdigheid' als zelfstandig naamwoord gebruikt en 'partijdig' als bijvoeglijk naamwoord.
Het begrip toets wordt hier beschouwd als een verbijzondering van het begrip test. Toets wordt gebruikt voor een meetprocedure van door onderwijs en studie verworven kennis, inzicht en vaardigheid op één of meer vakgebieden. Test wordt gebruikt voor een meetprocedure van niet door intentioneel onderwijs en studie verworven eigenschappen van de persoon (vgl. De Groot & Van Naerssen, 1969; Drenth, 1973; De Klerk, 1983).

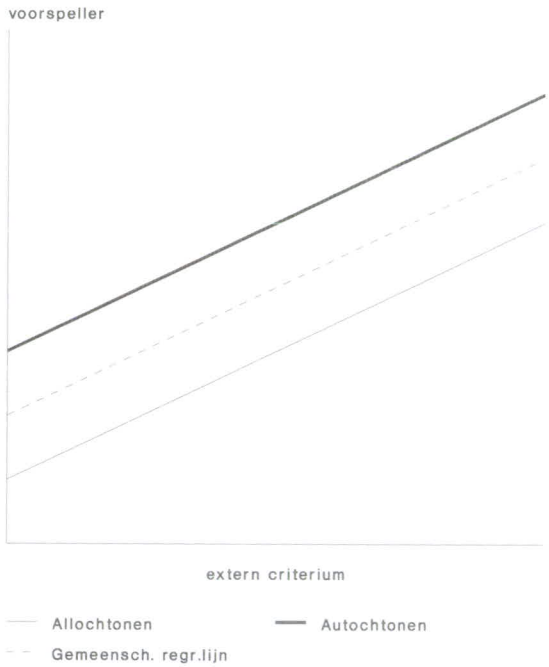
1.2.1 Onderzoek naar toetsbias

Toetsen worden in het algemeen ontwikkeld om voorspellingen te doen over buiten de toetssituatie liggend gedrag. Op basis van de behaalde toetsscore spreken we verwachtingen uit over feiten, waarvan we op zichzelf geen weet hebben, maar waarover we een conclusie formuleren op grond van eerder verworven kennis over de relatie tussen de toetsscore en de buiten de

toetssituatie liggende feiten (vgl. Drenth, 1973; De Klerk, 1983).
Wanneer de criteriumvaliditeit van een toets voor twee of meer subgroepen wordt onderzocht, spreken we van onderzoek naar toetsbias. Reynolds (1982) en Malpass & Poortinga (1986) definiëren toetsbias als het maken van systematische schattingsfouten bij het voorspellen van de positie op een extern criterium als een functie van een specifiek groepslidmaatschap. Jensen (1980: 381) zegt dat een toets partijdig is wanneer de hellingen, de intercepts en de schattingsfouten van de regressielijnen van twee subgroepen significant van elkaar verschillen. De 'American Educational Research Association' (AERA), de 'American Psychological Association' (APA) en de 'National Council on Measurement in Education' (NCME) onderschrijven in een gezamenlijke publicatie (1985) de opvatting van Jensen. Wanneer de regressielijnen van twee onderscheiden subgroepen samenvallen, dan voorspelt de toets het extern criterium voor beide groepen op dezelfde wijze. De intercepten en de hellingen van allochtone en autochtone leerlingen zijn gelijk en schattingsfouten in de predictie zijn niet gecorreleerd met groepslidmaatschap.
Een toets is partijdig wanneer de regressielijnen van onderscheiden subgroepen uit de populatie significant van elkaar verschillen en de gemeenschappelijke regressievergelijking gebruikt wordt om de positie van die subgroepen op het extern criterium te schatten. Deze situatie doet zich voor wanneer bij de predictie van het extern criterium geen onderscheid gemaakt wordt naar subgroepen.
Wanneer de regressielijnen niet samenvallen, kunnen zich drie situaties voordoen: de intercepten verschillen constant (a), de hellingen verschillen (b) en de intercepten en de hellingen verschillen (c) (Cronbach, 1972; Reynolds, 1982). Validiteit veronderstelt betrouwbaarheid. Bij de volgende situaties wordt ervan uitgegaan dat de meting voldoende betrouwbaar genoemd kan worden en dat alleen de predictieve validiteit voor subgroepen aan de orde is.

a *de intercepten verschillen significant*
Wanneer de intercepten verschillen en de hellingen niet, dan ontstaat er een situatie als in figuur 1.2.

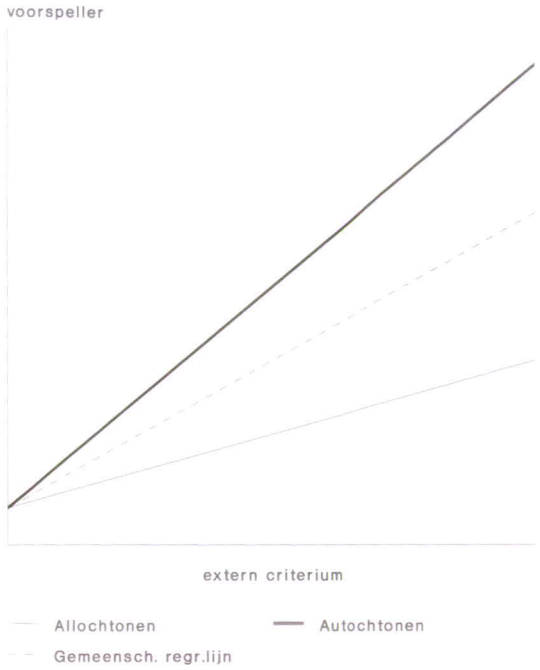
Figuur 1.2 De intercepten van de regressielijnen verschillen



Gebruik van de gemeenschappelijke regressievergelijking resulteert in bias ten nadele van de subgroep met de hoogste gemiddelde score op de toets (de voorspeller). Omdat de hellingen van beide subgroepen gelijk zijn, blijft de over- of onderschatting in de predictie constant en fluctueert niet als een functie van iemands score op de voorspeller. De mate van over- of onderschatting van het niveau op het extern criterium is dus onafhankelijk van iemands toetsscore. In figuur 1.2 leidt het gebruik van de gemeenschappelijke regressievergelijking tot overschatting van het criteriumniveau van allochtone leerlingen en tot onderschatting van het niveau van autochtone leerlingen.

b *de hellingen verschillen significant*
Figuur 1.3 geeft de situatie weer waarin de hellingen verschillen en de intercepten niet.

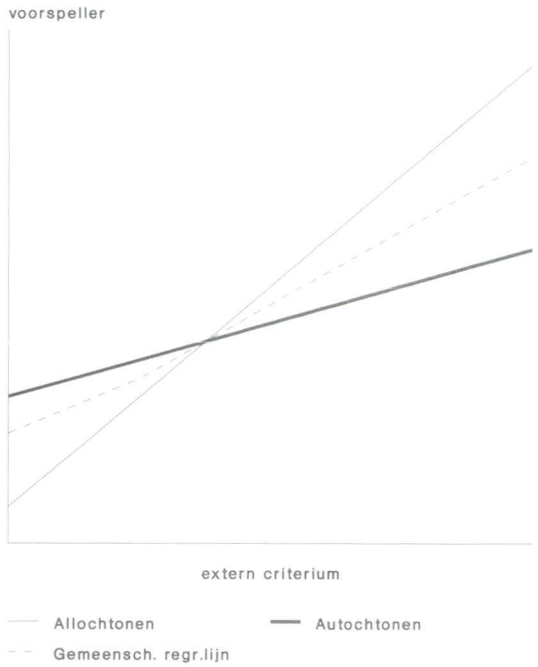
Figuur 1.3 De hellingen van de regressielijnen verschillen



In figuur 1.3 lopen de regressielijnen van de onderscheiden subgroepen niet parallel, hetgeen betekent dat bij gebruik van de gemeenschappelijke regressievergelijking het criteriumniveau van de subgroep met de hoogste gemiddelde toetsscore (voorspeller) onderschat wordt en dat het niveau van de groep met laagste score overschat wordt. De mate van toetsbias is hier niet bij elke toetsscore even groot, maar is afhankelijk van het scoreniveau op de voorspeller. Naarmate de toetsscore van allochtone leerlingen hoger is, wordt in figuur 1.3 het criteriumniveau sterker overschat; bij autochtone leerlingen gaat het om onderschatten.

c *de hellingen en de intercepten verschillen significant*
De situatie in figuur 1.4 is aanzienlijk complexer: zowel de hellingen als intercepten verschillen. Bij gebruik van de gemeenschappelijke regressievergelijking is de mate van toetsbias afhankelijk van het scoreniveau op de voorspeller, maar het scoreniveau op de voorspeller bepaalt ook of er sprake is van over- of onderschatting van het criteriumniveau. In figuur 1.4 wordt bij relatief lage toetsscores het criteriumniveau bij allochtone leerlingen overschat, bij relatief hoge scores onderschat. Alleen bij kruisende regressielijnen kan het voorkomen, dat het criteriumniveau van de subgroep met de laagste gemiddelde toetsscore wordt onderschat.

Figuur 1.4 De hellingen en de intercepten van de regressielijnen verschillen



Messick (1987) onderscheidt drie soorten beslissingen die met toetsscores genomen kunnen worden.

– *selectie*
Bij selectiebeslissingen wordt bepaald of een persoon wel of niet in aanmerking komt voor een bepaalde behandeling. Behandeling wordt hier gebruikt in ruime zin: het kan betrekking hebben op een aanvullend onderwijsprogramma, een therapeutische interventie of een arbeidsovereenkomst.

– *classificatie*
Bij classificatiebeslissingen worden alle personen over twee of meer soorten behandelingen verdeeld. De maatschappelijke waardering van de onderscheiden behandelingen is gelijk.

– *plaatsing*
Bij plaatsingsbeslissingen worden de leerlingen verdeeld over behandelingen die gezien de maatschappelijke waardering een rangorde vormen.

Messick (1987) vindt dat toetsscores die voor selectie, classificatie of plaatsing gebruikt worden, geëvalueerd moeten worden door longitudinaal onderzoek. De personen uit de onderscheiden subgroepen moeten gevolgd worden en na een bepaalde periode moeten gegevens verzameld worden over het criteriumgedrag. Met deze gegevens moet de predictieve validiteit van de scores per subgroep bepaald worden.

Cronbach (1972), Drenth (1972) en Messick (1987) zeggen dat het aangeven van de regressie van toetsscore op het extern criterium op zich onvoldoende is. Het is van belang om de determinanten van het criteriumgedrag te bepalen. Dit betekent in feite longitudinaal onderzoek met een model waarin een aantal relevante onafhankelijke variabelen, waaronder toetsscore, zijn opgenomen en waarin het criteriumgedrag als afhankelijke variabele fungeert.

Voor onderzoek naar de relatie toetsscore – extern criterium voor onderscheiden subgroepen is ook een longitudinaal model nodig. We moeten er immers rekening mee houden, dat de invloed van allerlei relevante variabelen op de relatie toetsscore – criterium bij elke onderscheiden subgroepen niet gelijk is. Het is uitermate belangrijk om vast te stellen welke factoren bij de onderscheiden subgroepen in dit verband differentiële effecten kunnen veroorzaken. Deze factoren moeten adequaat gemeten worden en vervolgens moeten de effecten van mogelijke determinanten van schoolloopbanen van de onderscheiden subgroepen in een longitudinaal model geschat worden.

Cronbach (1972), Jensen (1980), Reynolds (1982), Kok (1988) en Van de Vijver, Willemse & Van de Rijt (1993) merken op dat bij onderzoek naar toetsbias wordt aangenomen dat van het extern criterium een betrouwbare en valide operationalisatie beschikbaar is. Jensen (1980) en Van de Vijver, Willemse & Van de Rijt (1993) erkennen dat een onpartijdig extern criterium niet altijd voorhanden is, met name niet wanneer het extern criterium gebaseerd is op subjectieve, invalide observaties zoals bijvoorbeeld schoolcijfers. Wanneer het extern criterium ter discussie staat, kan volgens Jensen (1980) en Reynolds (1982) de aandacht beter uitgaan naar de constructvaliditeit van het meetinstrument, want uitspraken over de criteriumvaliditeit zijn dan eigenlijk niet mogelijk en niet toegestaan.

In Nederlands schoolloopbaanonderzoek wordt meestal het niveau dat een leerling na een bepaalde periode in het voortgezet onderwijs bereikt heeft als criterium voor schoolsucces gehanteerd. Uitgangspunt hierbij is dat de bereikte onderwijsposities verticaal (leerjaren) en horizontaal (van IBO tot VWO) verschillen in niveau en op één schaal gebracht kunnen worden. Vervolgens kan de regressie van de onafhankelijke variabelen (bijvoorbeeld: advies basisschool en toetsscore) op de schaal voor schoolsucces bepaald worden. Er blijken verschillende manieren te zijn om de bereikte onderwijsniveaus te schalen (Cremers, 1980; Tesser, 1986, Bosker, 1990, Uiterwijk, 1990b; Van der Velden, 1991). In verband met veranderingen in het voortgezet onderwijs moeten we er ook rekening mee houden dat een schaal voor bereikt onderwijsniveau een beperkte geldigheidsduur bezit.
Voor onderzoek naar toetsbias kan men als extern criterium een schaal voor bereikt onderwijsniveau construeren, maar bij de verantwoording van deze schaal moet ook aangegeven worden of deze schaal zelf onpartijdig is met betrekking tot de onderscheiden subgroepen.
Jungbluth, Van Langen & Vierke (1990: 91) stellen dat bij de overgang van basisonderwijs naar voortgezet onderwijs achteraf moeilijk vastgesteld kan worden of het advies van de basisschool of een toetsscore correct is geweest. Volgens hen heeft het advies van de basisschool een 'zich-zelf-waarmakend karakter'. Leerkrachtengedrag en het zelfbeeld van de leerling zullen leiden tot

zodanige verwachtingen dat het advies basisschool in de regel zijn eigen correctheid bevordert. Bovendien zullen 'systeemimmanente processen' met name categoriale scholen voor voortgezet onderwijs afhouden van op- en afstroom van leerlingen. Het is derhalve niet ondenkbaar dat bij het besluit om een leerling een andere school te adviseren wellicht onbedoeld meer factoren meespelen dan alleen de capaciteiten en het prestatieniveau van de leerling. Maar ook bij brede scholengemeenschappen kan men zich afvragen waar de selectie tijdens de brugperiode op gebaseerd is. Brede scholengemeenschappen zijn voor onderzoek naar toetsbias voor allochtone leerlingen belangrijk, omdat zij op deze schooltypen vergeleken met autochtone leerlingen oververtegenwoordigd zijn (Uiterwijk, 1990a). Tot nu toe is uit onderzoek weinig bekend over de vraag hoe op deze scholen selectieprocessen tot stand komen. Wijnstra (1984b), De Jong (1987), Uiterwijk (1990b), Driessen (1991a), Van Langen & Jungbluth (1992) en Meijnen & Riemersma (1992) constateren dat kinderen uit etnische minderheidsgroepen aan het einde van de basisschool gemiddeld een hoger advies krijgen dan de autochtone leerlingen met een vergelijkbare test- c.q. toetsscore. Het is niet uitgesloten dat leerkrachten van brede scholengemeenschappen net als hun collega's uit het basisonderwijs bij plaatsings- en overgangsbeslissingen het prestatieniveau van bepaalde subgroepen over- of onderwaarderen. Hierdoor is het vinden van een onpartijdig extern criterium een probleem.

De conclusie moet zijn dat het strikt genomen in de Nederlandse situatie onmogelijk is om te beoordelen of er bij een bepaalde toets sprake is van toetsbias vanwege het ontbreken van een onpartijdig extern criterium. Aan de andere kant moeten we vaststellen dat in de onderwijspraktijk toetsen en het advies basisschool een functie vervullen bij de schoolkeuze en de toelating tot het voortgezet onderwijs. Daardoor functioneert in de praktijk het bereikte onderwijsniveau wel degelijk als maat voor schoolsucces. Zo zeggen we bijvoorbeeld dat het advies van de basisschool goed is geweest, wanneer een leerling met een VWO-advies zonder doubleren in de derde klas VWO terecht komt. We zeggen echter ook dat de toetsuitslag onjuist was, wanneer een leerling met een score net onder het gemiddelde zonder doubleren eveneens in de derde klas VWO terecht komt. Uiteraard is het mogelijk dat het 'zich-zelf-waarmakend karakter' van het advies basisschool andere effecten heeft op de schoolloopbaan dan dat van de toetsuitslag, waardoor het moeilijk is om over juiste en onjuiste adviezen en scores te spreken.
Voor de onderwijspraktijk kan het evenwel van belang zijn te weten bij welke van onderscheiden subgroepen het advies basisschool hoger correleert met een schaal voor schoolsucces, bij welke subgroepen de toetsscore hoger correleert en bij welke subgroepen de correlaties van advies en score vergelijkbaar zijn. In deze studie wordt geen onderzoek naar toetsbias gedaan door aan de hand van het verschil tussen de regressielijnen van Eindtoetsscore op extern criterium van allochtone en autochtone leerlingen te bepalen of er bij de Eindtoets Basisonderwijs sprake is van toetsbias. Het ontbreken van een onpartijdig extern criterium maakt het in feite onmogelijk om te beoordelen of er bij de Eindtoets Basisonderwijs al dan niet sprake is toetsbias. Onderzoek naar toetsbias wordt hier opgevat als het nagaan van de predictieve validiteit van de Eindtoets Basisonderwijs voor onderscheiden etnische groepen in vergelijking met die van het advies van de basisschool.

1.2.2 Onderzoek naar itembias

Bij het ontwikkelen van een toets wordt een reeks items geconstrueerd die samen geacht worden een bepaald construct te representeren. De afzonderlijke items zijn operationalisaties van het construct dat de toets als geheel meet. In onderzoek naar de constructvaliditeit kan nagegaan worden of de items het construct voor onderscheiden subgroepen op vergelijkbare wijze representeren. Reynolds (1982) en Shepard (1982) stellen dat een item partijdig is, wanneer een toets bij de ene groep een ander construct meet dan bij de andere of wanneer de toets bij twee subgroepen wel hetzelfde meet maar dat niet met dezelfde nauwkeurigheid doet. Holland & Thayer (1986) zeggen kortweg dat partijdige items voor de ene subgroep een andere functie hebben dan voor de andere. In de Verenigde Staten wordt in plaats van over 'itembias' ook wel gesproken over 'Differential Item Functioning' (DIF).
Over de definitie van itembias blijken de meningen overeen te stemmen. Een item is partijdig wanneer leerlingen uit onderscheiden subgroepen, maar met een gelijke vaardigheid, een ongelijke kans hebben om het item goed te beantwoorden (Ironson, 1982; Angoff, 1982; Scheuneman, 1988; Verhelst, 1988; Kok, 1988; Hambleton & Rogers, 1989; Mellenbergh, 1989; Glas, 1991; Van de Vijver, 1991; Bügel, 1991; Glas & Ouborg, 1993). Als alle items van een toets het te meten domein (bijvoorbeeld het rekendomein 'kommagetallen') adequaat representeren, dan hebben leerlingen die even vaardig zijn in dat domein, een gelijke kans om een bepaald item uit die toets goed te beantwoorden. Van belang is dat voor het juist beantwoorden van de items een bepaalde populatie (bijvoorbeeld autochtone leerlingen) geen andere vaardigheden nodig heeft dan de vaardigheid die de items beogen te meten. De items meten in die populatie dan een eendimensionele vaardigheid ('kommagetallen'). Verder is van belang dat de leerlingen geclassificeerd kunnen worden naar de vaardigheid die de te onderzoeken items beogen te meten. Er moet dus een criterium beschikbaar zijn, waarmee de leerlingen van een bepaalde populatie (bijvoorbeeld autochtone leerlingen) ingedeeld kunnen worden in niveaugroepen. Dit criterium moet hetzelfde construct meten ('kommagetallen') als de te onderzoeken items pretenderen te meten. Vervolgens kan met statistische procedures onderzocht worden of leerlingen uit onderscheiden subgroepen (bijvoorbeeld autochtone en allochtone leerlingen), maar met een vergelijkbaar vaardigheidsniveau, een ongelijke kans hebben om het item goed te beantwoorden.
Holland & Thayer (1986) zeggen dat als resultaat van het classificeren van de leerlingen deze vergelijkbaar moeten zijn ten aanzien van
- het construct dat het item meet;
- het ontvangen onderwijsaanbod of andere relevante ervaringen;
- lidmaatschap van andere groepen.

Zij erkennen dat in de praktijk vrijwel altijd met minder genoegen moet worden genomen, hetgeen de trefzekerheid beperkt waarmee uitspraken over itembias gedaan kunnen worden.

Voor onderzoek naar itembias zijn verschillende statistische procedures beschikbaar, die in twee groepen verdeeld kunnen worden: klassieke testtheorie en itemresponsetheorie (vgl. Ironson, 1982; Kok, 1988; Hambleton & Rogers, 1989; Van de Vijver, 1991; Bügel & Glas, 1991; Glas & Ouborg, 1993).

a *Klassieke Testtheorie*
Klassieke testtheorieprocedures gaan van de aanname uit dat het totaal aantal goed gemaakte opgaven een goede schatting is van de te meten vaardigheid. Omdat deze aanname niet statistisch getoetst wordt, is een procedure gebaseerd op de klassieke testtheorie methodologisch eenvoudiger dan een itemresponse-theorie-procedure. De laatste jaren is de meest gebruikte klassieke testtheorie-procedure de Mantel-Haenszel-techniek (Holland & Wainer, 1993; Glas & Ouborg, 1993). Hierbij worden aan de hand van de totaalscore de leerlingen uit de onderscheiden subgroepen (bijvoorbeeld allochtone en autochtone leerlingen) ingedeeld in niveaugroepen. Vervolgens wordt de hypothese getoetst dat binnen deze niveaugroepen de p-waarde, het percentage leerlingen dat het item goed maakt, van het item bij allochtone en autochtone leerlingen gelijk is (Verhelst, 1988). Het classificeren naar niveaugroepen aan de hand van de totaalscore kan een probleem zijn, omdat de totaalscore ook de responsen op partijdige items kan bevatten. Hiervoor kan een oplossing gevonden worden door de totaalscore met behulp van een iteratieve procedure te 'zuiveren' van partijdige items. Eerst wordt een Mantel-Haenszel-analyse uitgevoerd waarbij alle items van de toets in kwestie zijn opgenomen in de totaalscore. Vervolgens worden de items die in de eerste analyse partijdig bleken te zijn, in de tweede analyse niet opgenomen in de totaalscore. Het is mogelijk dat er in de tweede analyse nieuwe partijdige items bijkomen, maar het is eveneens mogelijk dat items niet meer partijdig zijn die in de eerste analyse wel partijdig waren. Het iteratieve proces gaat door totdat er een verzameling onpartijdige items gevonden wordt waarop de totaalscore gebaseerd kan worden. Wanneer de leerlingen op basis van de 'gezuiverde' totaalscore zijn ingedeeld in niveau-groepen, wordt vervolgens voor elk item uit de toets de hypothese getoetst dat binnen de niveaugroepen de p-waarde van het item voor de onderscheiden subgroepen gelijk is. Bij het 'zuiveren' van de totaalscore doet zich de vraag voor of de overgebleven items het construct nog voldoende dekken. Dit is de vraag naar de inhoudsvaliditeit van de overgebleven items. Wanneer de onderzoeker aannemelijk kan maken dat resterende items het domein voldoende representeren, dan beschikken we over een onpartijdige operationalisatie van het te meten construct.

b *Itemresponsetheorie*
Procedures die gebaseerd zijn op een model uit de itemresponsetheorie (IRT) gaan van de aanname uit dat de geobserveerde itemresponsen verklaard kunnen worden vanuit één onderliggende vaardigheid, de latente trek. Onder een IRT-model wordt statistisch getoetst of de items één latente trek vertegenwoordigen. Als het IRT-model past, meten de items een eendimensionele vaardigheid. De kans op een goed antwoord wordt dan beschreven als een functie van persoons- en itemparameters. Leerlingen met dezelfde score op de latente trek hebben een gelijke kans om een item goed te beantwoorden onafhankelijk van de populatie waartoe ze behoren. Een belangrijke aanname bij eendimensionaliteit is dat de waarschijnlijkheid dat de toetsdeelnemer een item goed beantwoordt, een monotoon stijgende functie van de latente trek is. De itemkarakteristieke curve (item characteristic curve of ICC) geeft de relatie weer tussen de eendimensionele vaardigheid en de kans om het item goed te beantwoorden. Onder een IRT-model is onderzoek naar itembias het bepalen of de parameters van de ICC's van de onderscheiden subgroepen significant van elkaar

verschillen (Skaggs & Lissitz, 1988; Kok, 1988; Hambleton & Rogers, 1989; Hills, 1989; Mellenbergh, 1989; Camilli & Smith, 1990; Bügel & Glas, 1991). Er worden meestal drie parameters gebruikt worden om de ICC te beschrijven:
- de moeilijkheidsparameter, die het vaardigheidsniveau aangeeft;
- de discriminatieparameter, die aangeeft in welke mate de kans op een goed antwoord stijgt, naarmate de vaardigheid toeneemt;
- de raadparameter, die de kans aangeeft dat de toetsdeelnemer het item goed beantwoordt door te raden.

Een IRT-model is dan ook meestal op één, twee of drie parameters gebaseerd. Glas & Verhelst (1993) en Shealy & Stout (1993) wijzen op multidimensionele IRT-modellen waarmee vastgesteld kan worden in welke mate elk item uit een toets een beroep doet op twee of meer latente vaardigheden, maar deze relatief nieuwe modellen zijn wiskundig ingewikkeld en de bruikbaarheid ervan voor onderzoek naar itembias is vooralsnog beperkt.

Onder een IRT-model is sprake van itembias wanneer de geobserveerde responsen van de onderscheiden subgroepen (bijvoorbeeld allochtone en autochtone leerlingen) niet vanuit één en dezelfde latente trek verklaard kunnen worden. Itembias wordt hier nagegaan door eerst de items voor één subgroep (bijvoorbeeld autochtone leerlingen) te schalen. De items die blijken te passen op een schaal representeren bij autochtone leerlingen dezelfde latente trek. Vervolgens wordt bepaald of dezelfde items ook een latente trek vertegenwoordigen bij allochtone en autochtone leerlingen (vgl. Mellenbergh, 1989; Bügel & Glas, 1991). De items die bij de beide subgroepen niet op deze schaal passen zijn partijdig. Voor allochtone leerlingen zijn er kennelijk additionele vaardigheden in het geding.

De vraag naar *de beste statistische procedure* laat zich niet eenvoudig beantwoorden. Omdat bij IRT-modellen onderzocht wordt of de items bij het model passen, is deze benadering vergeleken met klassieke testtheorie-procedures theoretisch superieur. IRT-modellen zijn echter wiskundig ingewikkeld en ze zijn volgens Kok (1988: 28) onbetrouwbaar bij kleine steekproeven. Bovendien is de IRT-benadering niet volledig bruikbaar wanneer blijkt dat één of meer in een toets opgenomen items voor een bepaalde populatie niet bij het model passen (Glas, 1991). Deze items moeten dan bij deze analyses buiten beschouwing blijven, hoewel ze vanwege dit kenmerk voor onderzoek naar itembias juist interessant zijn.
Klassieke testtheorieprocedures kunnen gebruikt worden bij relatief kleine steekproeven, leveren op het eerste gezicht goed interpreteerbare statistische toetsen, maar maken niet duidelijk of de items de te meten vaardigheid adequaat representeren (Kok, 1988). Intraprasert (1986) concludeert na vergelijking van vijf itembiasdetectieprocedures dat een aantal van 400 – 500 waarnemingen per steekproef bij elke methode tot betrouwbare resultaten leidt. Bij Educational Testing Service (ETS) in de Verenigde Staten geldt als regel dat voor alle statistische procedures bij voorkeur steekproeven van 500 waarnemingen per subgroep beschikbaar moeten zijn (Zieky, 1993). In 6.1.3 komen we op de steekproefomvang terug.
Zowel bij IRT-modellen als bij klassieke testtheorieprocedures moet vastgesteld worden of we te maken hebben met niet-uniforme itembias. Er is van niet-uniforme itembias sprake wanneer een item partijdig is bij

laagpresterende en niet bij hoogpresterende niveaugroepen of omgekeerd (Uiterwijk, 1990a).
Het is niet ongebruikelijk om voor het opsporen van partijdige items zowel een procedure gebaseerd op het IRT-model als een klassieke testtheorie-procedure te gebruiken (Skaggs & Lissitz, 1988; Hambleton & Rogers, 1989; Hills, 1989; Camilli & Smith, 1990; Bügel & Glas, 1991; Hambleton & Jones, 1992; Glas & Ouborg, 1993). Hierdoor wordt duidelijk in welke mate er overlap bestaat tussen de gehanteerde procedures. Zo vonden Hambleton & Rogers (1989) dat de Mantel-Haenszel-procedure en een op het IRT-model gebaseerde procedure in het aanwijzen van partijdige en onpartijdige items bij 75 tot 80 % van de items overeenstemden. Hills (1989) vermeldt dat verschillende itembiasdetectie-procedures niet volledig overeenstemmen bij het aanwijzen van partijdige items en geeft bovendien aan dat itembiasindices aanzienlijk verschillen wanneer dezelfde detectieprocedure wordt toegepast op verschillende a-selecte steekproeven uit een populatie. Bij 33 analyses met zowel de Mantel-Haenszel-techniek als met op een IRT-model gebaseerde techniek op afzonderlijke a-selecte steekproeven uit dezelfde populatie blijkt geen enkel item 33 keer partijdig te zijn. Slechts zeven items bleken 20 van de 33 keer partijdig; van de in totaal 92 items waren 13 items nooit partijdig.
Uit het bovenstaande blijkt dat het moeilijk is om vast te stellen of een item partijdig is of niet. Wel kan aangegeven worden in welke mate een item bij de verschillende procedures partijdig is: bij alle, bij een deel of nooit (vgl. Uiterwijk, 1990a).

Kok (1988: 6) onderscheidt bij onderzoek naar itembias twee fasen.
In de *detectiefase* worden met statistische procedures beslissingen genomen over de vraag of items wel of niet partijdig zijn.
In de *verklaringsfase* worden naar aanleiding van de geconstateerde statistische itembias en op grond van andere kennis en inzichten hypothesen geformuleerd over mogelijke oorzaken van itembias.
De hypothesen kunnen betrekking hebben op de eigenschappen van toets-deelnemers maar ook op kenmerken van items, die verantwoordelijk zijn voor itembias. Verklaringen voor itembias kunnen ook gevonden worden door experimenteel en correlationeel onderzoek. Scheuneman (1982; 1985), Scheuneman & Steinhaus (1987), Kok, (1988); Bügel & Robben-Willems (1989), De Jong & Vallen (1989), Uiterwijk (1990a) en Coenen & Vallen (1991), Bügel & Glas (1991) en Uiterwijk & Vallen (1991) proberen door de inhoud van partijdige items te analyseren de oorzaken van itembias te achterhalen. Door te zoeken naar overeenkomstige kenmerken van partijdige en onpartijdige items kunnen aanwijzingen verkregen worden over oorzaken van itembias. Deze aanwijzingen kunnen mogelijk een richtsnoer vormen voor toetsontwikkelaars. Volgens Scheuneman & Steinhaus (1987) is het zeer moeilijk om achteraf vast te stellen welk element uit een item verantwoordelijk is voor itembias. Bevindt de bron van itembias bij bijvoorbeeld een vierkeuze-item voor begrijpend lezen zich in de tekst waarover de vraag wordt gesteld, in de introducerende itemtekst, in de geformuleerde vraag of in de vier antwoordmogelijkheden?
In verband met deze onzekerheid hebben de eventuele conclusies uit de verklaringsfase een voorlopig karakter. Herhaald onderzoek en een uitgebreide hoeveelheid items zijn nodig om bevestiging te vinden voor bepaalde veronderstellingen. Tot nu toe zijn de resultaten met betrekking tot het

opsporen van de oorzaken van itembias bescheiden (Scheuneman & Steinhaus, 1987; Uiterwijk & Vallen, 1991).

Het analyseren van de inhoud van partijdige items stemt overeen met wat we in 1.1 inhoudsanalyse genoemd hebben. Inhoudsanalyse heeft betrekkelijk weinig betekenis wanneer we dit doen bij items waarvan we slechts vermoeden dat ze partijdig zijn. Aan de verklaringsfase moet de detectiefase vooraf gaan, omdat we dan op empirische basis kunnen aangeven welke items voor kinderen uit etnische minderheidsgroepen partijdig zijn.

1.3 Onderzoeksvragen

Dit onderzoek richt zich op drie onderdelen. Ten eerste (1.3.1) gaat het om het beschrijven van trends in de schoolresultaten van allochtone en autochtone leerlingen. Met schoolresultaten worden hier de toetsscores op de (onderdelen van de) Eindtoets Basisonderwijs bedoeld en de gegevens over de toelating tot en de doorstroming in het voortgezet onderwijs. De groep allochtone leerlingen wordt hier onderverdeeld in diverse etnische groepen. Ten tweede (1.3.2) richt het onderzoek zich op de vraag hoe hoog de voorspellende waarde van de Eindtoets Basisonderwijs is voor de onderscheiden etnische groepen in vergelijking met die van het advies van de basisschool. Ten derde (1.3.3) gaat het onderzoek in op de vraag welke items partijdig zijn voor allochtone of autochtone leerlingen en waarom dat het geval is.

Het onderhavige onderzoek beoogt onder meer informatie te verschaffen over de predictieve en constructvaliditeit van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen. Uit itembiasonderzoek is bekend dat het in het algemeen moeilijk is om met zekerheid vast te stellen waarom een item partijdig is voor allochtone of autochtone leerlingen. De onderzoeker kan het beste voordat de items met statistische technieken onderzocht worden, hypothesen formuleren over mogelijke oorzaken van itembias. Omdat een item in feite bestaat uit een aantal elementen is het niet altijd duidelijk welk element van een partijdig item de bias veroorzaakt. De conclusies die na de inhoudsanalyse van de partijdige items getrokken worden, hebben derhalve nog een voorlopig karakter. Deze voorlopige conclusies zijn te beschouwen als hypothesen voor een eventuele volgende fase van het onderzoek. Voor itembiasonderzoek is een groot aantal items nodig en het onderzoek verloopt in feite via een aantal cycli. In verband hiermee zijn in dit onderzoek van de Eindtoetsdeelnemers uit 1987 en uit 1989 achtergrondgegevens verzameld. Door zowel in 1987 als in 1989 gegevens over de aan de Eindtoets Basisonderwijs te verzamelen is het ook mogelijk om onderzoeksgegevens van verschillende jaren te vergelijken. Met de data uit 1987 en 1989 worden trends in enerzijds toetsresultaten en anderzijds in toelatings- en doorstroomgegevens getraceerd en krijgen we een indruk van de stabiliteit van de predictieve validiteit van de Eindtoets Basisonderwijs en van het advies basisschool voor allochtone en autochtone leerlingen.

1.3.1 Trends in de schoolresultaten van allochtone en autochtone leerlingen

Met trends in schoolprestaties worden hier toetsresultaten op de (onderdelen van de) Eindtoets Basisonderwijs bedoeld en de toelatings- en doorstroomgegevens van deze leerlingen in het voortgezet onderwijs. De mogelijkheden van vergelijking van toetsresultaten zijn relatief groot, omdat de variabelen in beide jaren op sterk vergelijkbare wijze kunnen worden geoperationaliseerd. Bij vergelijking van cross-sectioneel onderzoek blijkt dit niet altijd het geval te zijn. Toetsen die hetzelfde construct pretenderen te meten, zijn niet altijd vergelijkbaar naar inhoud. De taaltoets die gebruikt wordt bij de Landelijke Evaluatie van het Onderwijsvoorrangsbeleid (Van Bergen, 1989) kent een andere opzet dan de taaltoets van Driessen (1990), die samengesteld is uit een beperkt aantal items van de Cito-Entreetoets. De taaltoets die Reezigt & Weide (1990) gebruiken, bestaat uit het taalonderdeel van de Cito-Eindtoets Basisonderwijs 1989. Zo bevat de taaltoets van Driessen opgaven over ontleden, terwijl dit domein ontbreekt in de taaltoets van Reezigt & Weide. In de taaltoets die gebruikt wordt voor de Landelijke Evaluatie van het Onderwijsvoorrangsbeleid ontbreken spelling en ontleden. Het taalonderdeel uit deze laatste toets bestaat uit tweekeuze-opgaven, de toetsen van Driessen en Reezigt & Weide bestaan uit vierkeuze-opgaven. Waarschijnlijk zullen de genoemde taaltoetsen in aanzienlijke mate correleren, maar er blijft ruimte voor onverklaarde variantie.

Van de Eindtoets Basisonderwijs 1987 en 1989 zijn de totaalscores (zie Hoofdstuk 3: de geëquivaleerde standaardscores) volledig vergelijkbaar, maar dit geldt ook in sterke mate voor de scores van de toetsonderdelen Taal, Rekenen en Informatieverwerking. De toetsspecificatie is voor beide jaren gelijk, elk toetsonderdeel bestaat in beide jaren uit 60 opgaven en de verdeling van de opgaven over de verschillende opgavenrubrieken (zie Hoofdstuk 3) komt sterk overeen. In welke mate trends in totaal- en onderdeelscores informatief zijn, weten we echter niet. Duidelijk is nog niet in hoeverre de items van de Eindtoets het te meten construct bij allochtone leerlingen valide meten. Het is niet duidelijk in hoeverre stijging of daling van scores corresponderen met toe- of afname van de vaardigheid in het te meten construct. Met de interpretatie van de scores is vooralsnog voorzichtigheid geboden.
Over de representativiteit van de toetsdeelnemers uit 1987 en 1989 voor alle leerlingen in groep acht van het basisonderwijs in die jaren valt weinig te zeggen. De toetsdeelnemers zijn de leerlingen van de scholen die deelnamen aan de Eindtoets Basisonderwijs van dat jaar (zie inleiding Hoofdstuk 4 en 4.1). Er ontbreken echter mogelijkheden om het prestatieniveau van de toetsdeelnemers uit 1987 en 1989 te vergelijken met het niveau van alle leerlingen uit het laatste leerjaar van het basisonderwijs. Dit betekent uiteraard ook dat we niet weten in hoeverre de autochtone leerlingen en de leerlingen uit de verschillende etnische minderheidsgroepen die in de betreffende jaren aan de toets deelnamen, representatief zijn voor al hun groepsgenoten in groep acht van het basisonderwijs.
De representativiteit van leerlingen uit de verschillende etnische minderheidsgroepen is overigens ook een probleem wanneer een steekproef uit alle basisscholen getrokken zou worden. Het verdient aanbeveling om een gestratificeerde steekproef te trekken wanneer de kans niet bij elke school gelijk

is dat de te onderzoeken subpopulatie in groep acht van een school aanwezig is. Hiervoor is informatie nodig over de spreiding van de onderscheiden subpopulatie(s) over de scholen. In het onderzoek van Driessen (1990: 71) bestaat de populatie uit de leerlingen van groep acht van basisscholen met minimaal één allochtone leerling in die jaargroep. Driessen zegt dat over de representativiteit van zijn onderzoek weinig valt te zeggen, vanwege het ontbreken van toegankelijke overzichtsstatistieken van relevante variabelen. Dit is op zich juist, maar men kan wel betwijfelen of de autochtone leerlingen in klassen met allochtone leerlingen representatief zijn voor alle autochtone leerlingen in Nederland.

De vergelijkbaarheid van de gegevens van leerlingen uit etnische minderheden en autochtone leerlingen wordt ook beperkt door de verschillende indelingscriteria die onderzoekers ten aanzien van etniciteit hanteren. Tesser & Mulder (1990) spreken van een Turkse leerling wanneer de vader in Turkije geboren is en wanneer de wegingsfactor voor de formatieregeling Wet op het Basisonderwijs 1.90 is. Reezigt & Weide (1990) spreken daarentegen van een Turkse leerling wanneer de moeder in Turkije geboren is. De Jong, Uiterwijk, Kerkhoff & Vallen (1987) spreken pas van een Turkse leerling wanneer het herkomstland van de beide ouders Turkije is (bij één-ouder-gezinnen geldt het herkomstland van de ouder/verzorger bij wie het kind woont). Ook bij de Landelijke Evaluatie van het Onderwijsvoorrangsbeleid moeten beide ouders in het herkomstland geboren zijn (Weide & Van der Werf, 1990). Omdat uit Van 't Hof & Dronkers (1992) blijkt dat de kinderen uit etnisch homogene gezinnen een hoger advies van de basisschool krijgen en een hogere positie in het voortgezet onderwijs innemen dan kinderen uit etnisch heterogene gezinnen, kunnen deze verschillende operationalisaties van etniciteit wellicht invloed uitoefenen op de uitkomsten van een onderzoek.
Door de toetsinhoud en de operationalisaties van allochtone en autochtone leerlingen in twee metingen vergelijkbaar te houden, is het mogelijk om de eerste onderzoeksvraag als volgt te formuleren:

1 *Hoe ontwikkelen de Eindtoetsscores van allochtone en autochtone leerlingen zich van 1987 tot 1989?*

De verdeling van leerlingen over de onderwijstypen in het voortgezet onderwijs is reeds lang onderwerp van onderzoek. Vooral onderwijssociologisch onderzoek richt zijn aandacht op deze onderwijsperiode gezien het belang van de hier gemaakte keuze voor het vervolg van de schoolloopbaan (Meijnen, 1979; Tesser, 1986; Bosker, 1990). Eerst in de tachtiger jaren komen landelijke onderzoeksresultaten beschikbaar over de instroom van allochtone leerlingen in het voortgezet onderwijs gekoppeld aan relevante achtergrondvariabelen (Van Esch, 1983; Wijnstra, 1984b). Daarna zijn er meer onderzoeken op landelijk niveau gevolgd (Driessen, 1990; Tesser, Mulder & Van der Werf, 1991; Mulder & Tesser, 1991; Mulder, 1993). Er is in feite pas een begin gemaakt met het vinden van empirisch gefundeerde verklaringen voor de verschillende instroom maar ook voor de verschillen in schoolsucces van allochtone en autochtone leerlingen in het voortgezet onderwijs (vgl. Meijnen & Riemersma, 1992).
Het voortgezet onderwijs verkeert niet in een status quo. Door fusie van categoriale scholen tot brede scholengemeenschappen, door veranderingen in

de aantrekkelijkheid van bepaalde schooltypen, door veranderingen in het onderwijsaanbod kan de instroom van verschillende schooltypen wijzigen. In dit verband kunnen ontwikkelingen in de instroom van allochtone en autochtone leerlingen relevante informatie leveren. De tweede onderzoeksvraag kan als volgt geformuleerd worden:

2 *Hoe verloopt de toelating- en doorstroming van allochtone en autochtone leerlingen in het voortgezet onderwijs?*

De eerste en tweede onderzoeksvraag komen in hoofdstuk vier aan bod.

1.3.2 De predictieve validiteit van de Eindtoets Basisonderwijs voor de onderscheiden etnische groepen in vergelijking met die van het advies van de basisschool

De onderwijsposities die leerlingen na een bepaalde periode in het voortgezet onderwijs innemen, zijn schaalbaar. Deze schaal, die als indicator voor schoolsucces is te beschouwen, kan in een longitudinaal model als afhankelijke variabele functioneren met als voorspeller bijvoorbeeld toetsscores en advies van het basisonderwijs. De relaties in dit model kunnen mogelijk verklaringen aanreiken over de verschillende instroom van leerlingen uit diverse etnische minderheidsgroepen en autochtone leerlingen in het voortgezet onderwijs, hetgeen mogelijk van belang is voor de schoolkeuze- en toelatingspraktijk van allochtone leerlingen. De derde onderzoeksvraag is dan ook als volgt geformuleerd:

3 *Hoe hoog is voor allochtone en autochtone leerlingen de voorspellende waarde van de Eindtoets Basisonderwijs in vergelijking met het advies van de basisschool?*

Deze derde onderzoeksvraag komt aan de orde in hoofdstuk vijf.

1.3.3 Itembias voor allochtone leerlingen

In onderzoek naar itembias kunnen twee complementaire fasen worden onderscheiden. In de eerste fase, die Kok (1988) de detectiefase noemt, worden met statistische procedures partijdige items opgespoord. In de tweede fase, de verklaringsfase (Kok, 1988), wordt ingegaan op de vraag wat bij een bepaald item de oorzaak van itembias zou kunnen zijn.
Eerst wordt de detectiefase aan de orde gesteld. Uit de definitie van itembias (zie par 1.2.2) volgt, dat nagegaan moet worden of de leerlingen uit onderscheiden subgroepen maar met dezelfde vaardigheid, een gelijke kans hebben om een item goed te beantwoorden. Er moet dus een procedure gehanteerd worden waarmee leerlingen geclassificeerd worden naar niveaugroepen en daarna kan statistisch getoetst worden of de itemresponsen van die niveaugroepen significant verschillen. De keuze van deze procedure is niet zonder problemen. Uit de literatuur blijkt dat de verschillende statistische itembiasdetectieprocedures, toegepast op dezelfde toets, niet tot dezelfde resultaten leiden (Intraprasert, 1986; Skaggs & Lissitz, 1988; Hambleton & Rogers, 1989; Hills, 1989; Camilli & Smith, 1990; Bügel & Glas, 1991; Hambleton & Jones, 1992).

De verschillende resultaten kunnen toegeschreven worden aan een aantal oorzaken. Centraal staat evenwel de eendimensionaliteit van de toets. Wanneer de responsen op de items van een toets volledig vanuit één latente trek te verklaren zijn, dan zullen de resultaten van de verschillende itembiasdetectie-procedures en van de verschillende steekproeven aanzienlijk overeenstemmen. Wanneer niet statistisch getoetst wordt of de items een eendimensionele vaardigheid representeren, is het niet duidelijk aan de hand van welke vaardigheid leerlingen geclassificeerd worden naar niveaugroepen. Toetsen die een breed domein bestrijken (bijvoorbeeld: rekenen, taal) zullen in dit verband meer problemen opleveren dan toetsen over meer specifieke domeinen (bijvoorbeeld: basiskennis van natuurlijke en decimale getallen, spelling van de werkwoordsvormen).
De Eindtoets Basisonderwijs is te beschouwen als een toets die een aantal brede domeinen bestrijkt. De toets- en itemanalyses worden gebaseerd op de klassieke testtheorie (vgl. Engelen & Uiterwijk, 1990; Uiterwijk & Engelen, 1992) en vanuit dat oogpunt ligt het voor de hand om gebruik te maken van de Mantel-Haenszel-procedure. Omdat de keuze van de statistische procedure van invloed kan zijn op de vraag welke items partijdig zijn, is het noodzakelijk om de resultaten van minstens twee procedures te vergelijken. Omdat het met een procedure die gebaseerd op een IRT-model mogelijk is statistisch te toetsen welke items bij het model passen, gaat voor de tweede procedure de voorkeur uit naar een IRT-model.
De vierde en vijfde onderzoeksvraag zijn als volgt geformuleerd:

4 *Welke statistische procedure verdient de voorkeur voor het opsporen van itembias bij de Eindtoets Basisonderwijs?*

5 *Welke opgaven zijn voor allochtone leerlingen significant moeilijker of makkelijker dan voor autochtone leerlingen met een vergelijkbaar prestatieniveau?*

Deze twee onderzoeksvragen komen in hoofdstuk zes aan de orde.

In de detectiefase van onderzoek naar itembias worden met statistische procedures partijdige items opgespoord (onderzoeksvraag 5). In de tweede fase (verklaringsfase) wordt ingegaan op de vraag wat bij een bepaald item of een reeks van items de oorzaak van itembias zou kunnen zijn. Met het zoeken naar oorzaken van itembias voor allochtone leerlingen, is niet alleen in Nederland maar ook in andere landen, bijzonder weinig ervaring opgedaan. Volgens Schmitt, Holland & Dorans (1992) zijn er voor de geringe ontwikkeling inzake het vinden van oorzaken van itembias drie redenen aan te wijzen.
In de eerste plaats is onderzoek naar itembias relatief nieuw en tot nu toe is de meeste aandacht uitgegaan naar de ontwikkeling van statistische procedures voor het detecteren van partijdige items. In de tweede plaats veronderstelt het achterhalen van oorzaken van itembias voor allochtone leerlingen een theorie over de vraag waarom items voor de onderscheiden etnische groepen moeilijk zijn. Maar omdat de etnische groepen intern vaak in veel opzichten heterogeen zijn, kunnen de verschillen tussen de etnische groepen moeilijk beschreven worden. In de derde plaats is het opsporen van oorzaken van itembias complex, omdat bij een bepaald item verschillende oorzaken een rol kunnen spelen.

Scheuneman (1985) is van mening dat voor het goed beantwoorden van items, ook items die een eendimensionele vaardigheid meten, verschillende samenhangende (deel)vaardigheden nodig zijn en dat het achteraf moeilijk is om vast te stellen welk element van een item verantwoordelijk is voor itembias. Scheuneman & Steinhaus (1987) zijn van mening dat voor de statistisch toetsing per item hypothesen geformuleerd moeten worden over oorzaken van bias. Omdat goed gefundeerde taalkundig-inhoudelijke verklaringen voor itembias voor allochtone leerlingen geheel ontbreken, kunnen voor het formuleren van hypothesen over potentiële bronnen van itembias voor leerlingen uit etnische minderheidsgroepen onderzoeksgegevens over verschillen tussen schoolprestaties van allochtone en autochtone leerlingen een kader aanreiken. Bovendien kunnen sociaal-culturele verschillen tussen allochtone en autochtone leerlingen van belang zijn in zoverre ze een potentiële bron van itembias zijn. Duidelijk moet gesteld worden dat verschil in schoolprestaties van onderscheiden groepen niet hetzelfde is als bias. In 1.2 is reeds aangegeven waarom 'moeilijkheid' en 'itembias' geen identieke concepten zijn. Bij het formuleren van hypothesen over itembias moet men wel te rade gaan bij optredende verschillen in schoolprestaties. Of verschillende prestaties daadwerkelijk tot itembias leiden, moet in de detectiefase statistisch getoetst worden. In 2.2 wordt ingegaan op potentiële bronnen van itembias.

In het onderhavige onderzoeksproject zijn de items die in de detectiefase partijdig bleken te zijn geanalyseerd met het oog op de vraag welk element van een bepaald item de oorzaak van itembias zou kunnen zijn. Bij het achterhalen van mogelijke oorzaken van itembias werden drie groepen personen betrokken. Eerst hebben de projectmedewerkers (van KUB en Cito) onafhankelijk van elkaar nagegaan welk element van een item verantwoordelijk zou kunnen zijn voor bias. Hierbij hebben de potentiële bronnen van itembias (2.2) als hypothesen gefungeerd. De overeenkomsten in de analyses van de projectmedewerkers zijn geïnventariseerd en beschreven (7.1). Vervolgens is een aantal niet bij het onderzoeksproject betrokken experts gevraagd aan te geven of de aan hen voorgelegde items moeilijker dan wel makkelijker voor allochtone leerlingen zijn en welke oorzaken daarvoor te geven zijn. Verder zijn bij het onderzoeksproject ook allochtone en autochtone leerlingen uit groep acht van het basisonderwijs betrokken. Met een kleinschalig hardop-denken-experiment is nagegaan hoe vaak allochtone en autochtone leerlingen bij partijdige items een fout antwoord geven ten gevolge van een element uit het item dat voorlopig als bron van itembias is aangewezen. Daarnaast is onderzocht hoe vaak bij gemanipuleerde items door de itemmanipulatie een goed antwoord wordt gegeven. Gemanipuleerde items zijn items waarbij het itemelement dat voorlopig als biasbron is aangewezen, vervangen is door een itemelement waarvan verwacht wordt dat het geen bias veroorzaakt. In 7.3 wordt verslag gedaan van het hardop-denken-experiment waarin leerlingen uit groep acht van het basisonderwijs aangeven hoe ze de oorspronkelijke (partijdige), respectievelijk de gemanipuleerde items hebben opgelost.
De zesde onderzoeksvraag is als volgt geformuleerd:

6 *Welke bronnen van itembias voor allochtone leerlingen bevatten de opgaven van de Eindtoets Basisonderwijs 1987 en 1989.*

Op de laatste onderzoeksvraag wordt in hoofdstuk zeven ingegaan.

2 Potentiële bronnen van toets- en itembias

In 1.2.1 is aangegeven dat het onderzoek naar toetsbias in deze studie opgevat wordt als het nagaan van de predictieve validiteit van de Eindtoets Basisonderwijs voor de onderscheiden etnische groepen in vergelijking met die van het advies basisschool. In 2.1 wordt ingegaan op mogelijke oorzaken van eventuele verschillen in de voorspellende waarde van toets en advies ten aanzien van allochtone en autochtone leerlingen.
In 1.2 is gesteld dat itembias niet hetzelfde is als moeilijkheid. Bij het formuleren van hypothesen over itembias moet men wel te rade gaan bij verschillen in schoolresultaten. In 2.2 wordt ingegaan op mogelijke bronnen van itembias.

2.1 Mogelijke determinanten van verschillen in de predictieve validiteit van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen in vergelijking met het advies basisschool

Onderzoek om te beoordelen of er bij een bepaalde toets wel of niet sprake is van toetsbias, is in de Nederlandse situatie strikt genomen zeer moeilijk, vanwege het ontbreken van een onpartijdig extern criterium (zie 1.2.1). Onderzoek naar toetsbias wordt in deze studie opgevat als het vergelijken van de regressie van het advies basisschool en de Eindtoets Basisonderwijs op een schaal voor schoolsucces ten aanzien van allochtone en autochtone leerlingen. Hierbij moet ook rekening gehouden worden met andere variabelen die rechtstreeks of via advies en toetsscore van invloed kunnen zijn op schoolsucces.
Er zijn redenen om aan te nemen dat de predictieve validiteit van toets en advies kunnen verschillen.

Zoals in 1.2.1 is opgemerkt, is een toets partijdig wanneer de regressielijnen van onderscheiden subgroepen significant verschillen en de gemeenschappelijke regressievergelijking gebruikt wordt op de positie van elke subgroep op het extern criterium te schatten. Wanneer de regressielijnen elkaar niet kruisen, wordt het niveau op het extern criterium van de subgroep met de laagste gemiddelde score op de voorspeller overschat. Omgekeerd wordt het criteriumniveau van de subgroep met de hoogste gemiddelde score onderschat. Wanneer de regressielijnen van de subgroepen elkaar kruisen is de over- of onderschatting afhankelijk van de hoogte van de score op de voorspeller. Het gebruik van een gemeenschappelijke regressievergelijking bij verschillende gemiddelde scores op de voorspeller is dus een aanwijzing voor verschillen in de predictieve validiteit van toetsen.
Schoolvorderingentoetsen voor schoolkeuze meten idealiter een construct waarvan de relatie met het extern criterium is aangetoond. Deze relatie wordt gelegd door het schoolsucces van de toetsdeelnemers te onderzoeken. Omdat deze relatie nooit perfect is, zal de schatting van de positie in het voortgezet onderwijs op grond van de behaalde toetsscore altijd een zekere mate van onbetrouwbaarheid kennen. Gezien de veranderingen in het voortgezet onderwijs moeten deze toelatings- en doorstroomonderzoeken periodiek

herhaald worden, hetgeen niet bij alle toetsen en tests het geval is. Bij veranderingen in het voortgezet onderwijs kan bijvoorbeeld gedacht worden aan een daling of stijging van het gemiddelde prestatieniveau van de leerlingen in bepaalde onderwijstypen; aan veranderingen in het curriculum van een schooltype; aan veranderingen die gepaard gaan met het fuseren van categoriale scholen tot brede scholengemeenschappen en dergelijke.
Schoolvorderingentoetsen bedoeld voor de schoolkeuze meten de mate waarin de leerling bepaalde basisschoolleerstof verworven heeft en moeten tegelijkertijd relevante informatie voor de schoolkeuze verschaffen. Deze toetsen moeten zowel aan eisen voldoen in verband met de predictieve validiteit als aan eisen in verband met constructvaliditeit.
Verwacht mag worden dat toetsen het schoolsucces van allochtone leerlingen in het voortgezet onderwijs zullen overschatten, omdat bij toetsen over het algemeen een gemeenschappelijke regressievergelijking gebruikt wordt om de positie op het extern criterium te schatten. Aangezien allochtone leerlingen meestal lagere toetsscores hebben dan autochtone leerlingen zal een toets het schoolsucces van allochtone leerlingen overschatten en van autochtone leerlingen onderschatten (vgl. Figuur 1.2).

Het advies van de basisschool is een indicator voor het algemene prestatieniveau van de leerling in termen van de onderwijstypen van het voortgezet onderwijs. De achtjarige ervaring die het leerkrachtenteam met deze leerling heeft opgedaan, wordt in feite geprojecteerd op een schaal waar de schaalpunten gevormd worden door de schooltypen (van IBO tot VWO). Hier wordt niet de beoordeling van het algehele niveau van de leerling als zodanig als probleem opgevoerd, maar de transformatie ervan naar de schooltypenschaal. Het is gezien de hierboven genoemde veranderingen in het voortgezet onderwijs niet altijd duidelijk of de schooltypen nog steeds dezelfde eisen aan de leerlingen stellen. Over het algemeen krijgen leerkrachten uit de basisschool slechts zeer beperkte feedback over het schoolsucces van hun ex-leerlingen en kunnen ze dus hun schoolkeuze-adviezen nauwelijks evalueren. Daardoor is het mogelijk dat de realiteitswaarde van het door de basisschool gehanteerde externe criterium niet altijd optimaal is.

In het algemeen is er tot op heden relatief weinig onderzoek verricht waarin de effecten van determinanten van schoolloopbanen (onder andere: toets en advies basisschool) van allochtone en autochtone leerlingen in een longitudinaal model nauwkeurig zijn geschat (vgl. Driessen, 1990; Van Langen & Jungbluth, 1990; Jungbluth, Van Langen & Vierke, 1990; Mulder & Tesser, 1991). Ook onderzoek naar de relatie tussen schoolkeuze-adviezen, toetsscores en schoolkeuze van allochtone en autochtone leerlingen bestaat in Nederland nog niet zo lang. Wijnstra (1984b) vond als eerste dat allochtone leerlingen in Nederland hogere adviezen kregen dan op grond van hun toetsscores (Eindtoets Basisonderwijs 1982) mocht worden verwacht. De Jong (1987) constateerde dat allochtone leerlingen een hoger advies kregen dan gezien hun testscore (GALO) te verwachten viel. Uiterwijk (1990a), Driessen (1991a) en Van Langen & Jungbluth (1992) komen later tot dezelfde conclusie. Het is nog niet volledig duidelijk welke mechanismen achter deze overadvisering van allochtone leerlingen ten grondslag liggen. Het is denkbaar dat de directeuren uit het basisonderwijs de lagere beheersing van de Nederlandse taal als een

tijdelijk probleem zien en deze leerlingen bij de schoolkeuze eerder het voordeel van de twijfel geven dan autochtone leerlingen met een vergelijkbare beheersing van het Nederlands (vgl. Meijnen & Riemersma, 1992). Mulder & Tesser (1991) en Van Langen & Jungbluth (1992) constateren dat allochtone meer dan autochtone ouders het schoolkeuze-advies van de basisschool te laag vinden. Het is mogelijk dat dit van invloed is op het advies basisschool en op de uiteindelijke schoolkeuze.

Samengevat zijn de volgende mogelijke determinanten van verschillen in de predictieve validiteit van belang.
- Het gebruik van een gemeenschappelijke regressievergelijking om het schoolsucces in het voortgezet onderwijs te schatten bij subgroepen met verschillende gemiddelde scores op de voorspeller.
- De transformatie van de beoordeling van het niveau van een leerling (advies basisschool) naar een schaal voor schoolsucces.

Uit onderzoek volgt de verwachting dat het advies basisschool voor allochtone leerlingen een overschatting geeft van het schoolsucces in het voortgezet onderwijs. Verder wordt verwacht dat een toets het schoolsucces van allochtone leerlingen eveneens overschat en dat van autochtone leerlingen onderschat.

2.2 Mogelijke bronnen van itembias voor allochtone leerlingen

In de paragraaf over onderzoek naar itembias (1.2.2) wekken de begrippen 'construct', 'vaardigheid' en 'eendimensionaliteit' de indruk, dat het om enkelvoudige, goed omschreven psychologische vaardigheden gaat. Bij inhoudsanalyse van items, die een eendimensionele schaal vormen, wordt echter duidelijk dat het in feite om een complex van vaardigheden gaat. Items zijn operationalisaties van een construct en maken gebruik van contexten die ook een beroep doen op bepaalde (deel)vaardigheden. De volgende drie rekenitems geven een indruk van dergelijke (deel)vaardigheden. Ze maken samen met 14 andere items deel uit van de eendimensionele schaal 'Toepassingen procenten', die ontwikkeld is in het kader van de Periodieke Peiling van het Onderwijsniveau (Wijnstra, 1988).

Voorbeeld 1

Joop droomt dat hij zo veel geld op de bank heeft staan dat hij alleen van de rente kan leven. Hij zou dan *f* 3000,– per maand willen hebben. De bank geeft 6% rente per jaar. Hoeveel geld zou hij dan op de bank moeten hebben?

f _____

Voorbeeld 2

De heer Van Dam heeft een hypotheek van *f* 60 000,– tegen een rente van 8,4%. De rente daalt tot 7,9%.
Hoeveel gulden voordeel levert dit de heer Van Dam op per jaar?

f _____

Voorbeeld 3

Iemand koopt voor *f* 800,– aan spullen. Hij moet 30% van dit bedrag meteen betalen.
De rest mag hij in 10 gelijke delen betalen.
Hoe groot zijn die delen?

f _____

Voor de oplossing van deze drie opgaven moeten leerlingen in contexten percentageberekeningen uitvoeren, waarbij meer dan één oplossingsstrategie gevolgd kan worden. Bij elke opgave moet de leerling zich inleven in een bepaalde context en tegen de achtergrond van die context moet de leerling kiezen voor een bepaalde rekenoperatie. Elke opgave veronderstelt reken-, maar ook taalvaardigheden die recentelijk of langer geleden op de basisschool verworven zijn.
Bij de eerste opgave moeten de leerlingen weten dat *f* 3000,– een uitkomst is van een bepaalde bewerking. De leerlingen moeten vervolgens de bewerking reconstrueren, rekening houdend met de (in de lagere leerjaren aangeleerde) relatie maand – jaar. Bij de tweede opgave moet het verschil tussen beide rentepercentages in relatie gebracht worden met *f* 60 000,–. Bij de derde opgave moeten de leerlingen *f* 800,– op een bepaalde manier in twee ongelijke delen verdelen en daarna moeten zij het grootste deel weer in tien gelijke delen verdelen.

De leerlingen moeten in dit soort contexten voldoende aanwijzingen aantreffen voor de keuze van een adequate rekenoperatie. De contexten veronderstellen een gemeenschappelijk referentiekader. Het is niet altijd mogelijk om precies vast te stellen welke taal- en rekenvaardigheden hiervoor nodig zijn. Voor de beantwoording van de eerste opgave is het waarschijnlijk nodig dat de leerling weet wat 'van de rente leven' is en weet dat *f* 3000,– per maand de rente van een bepaalde hoofdsom is. Bij de tweede opgave is het niet nodig dat een leerling exact weet wat een hypotheek is. Wel wordt verondersteld dat alle toetsdeelnemers weten dat een hypotheek een schuld en geen bezit vertegenwoordigt. Toch weten we vooraf niet zeker in welke mate de kennis van het woord 'hypotheek' het goed beantwoorden van deze opgave faciliteert. Bij de derde opgave is de context door het begrip 'spullen' weinig specifiek. Bovendien is de formulering '... koopt voor *f* 800,– aan spullen' misschien niet voor alle toetsdeelnemers even bekend. Het is mogelijk dat de zin 'Iemand

koopt voor *f* 800,– een fiets.' tot andere toetsresultaten leidt.

Items die een eendimensionele schaal vormen, doen in feite een beroep op een aantal sterk samenhangende deelvaardigheden.
In geval van itembias meten items bij de ene subgroep wat anders dan bij de andere subgroep. Anders gezegd: bij de ene subgroep representeren de items een eendimensionele vaardigheid, bij de andere subgroep een multidimensionele. Bij partijdige items zijn er voor een subgroep kennelijk een aantal (deel)vaardigheden in het geding, waarvan er één of meer verantwoordelijk zijn voor multidimensionaliteit. In de verklaringsfase moet vastgesteld worden welke deelvaardigheden voor de partijdigheid van het item verantwoordelijk zijn (vgl. Kok, 1988).

Bij onderzoek naar de oorzaken van itembias is het moeilijk om met zekerheid vast te stellen waarom een bepaald item partijdig is voor bijvoorbeeld leerlingen uit etnische minderheidsgroepen (Scheuneman, 1985; Scheuneman & Steinhaus, 1987; De Jong & Vallen, 1989; Coenen & Vallen, 1991; Uiterwijk & Vallen, 1991). In verband met deze onzekerheid moet de onderzoeker idealiter, voordat met statistische technieken wordt bepaald welke items partijdig zijn, hypothesen formuleren over mogelijke oorzaken van itembias. Deze hypothesen kunnen geformuleerd worden op basis van hetgeen uit de literatuur bekend is over de factoren die verschillen tussen de onderwijsresultaten van de onderscheiden subgroepen verklaren. De rest van dit hoofdstuk biedt daarom een literatuuroverzicht dat beoogt richting te geven aan het formuleren van hypothesen over potentiële bronnen van itembias voor leerlingen uit etnische minderheidsgroepen. Het gaat daarbij om (deel)vaardigheden, waarvan uit onderzoek is gebleken of waarvan te verwachten is dat allochtone en autochtone leerlingen aan het einde van het basisonderwijs deze in verschillende mate beheersen. Het gaat dus om moeilijkheid en niet om bias. Verschillen in moeilijkheidsgraad zijn in het kader van dit onderzoek echter van belang, omdat ze *potentiële* bronnen van bias zijn.

De Jong & Vallen (1989) stellen dat specifieke bronnen van itembias gezocht kunnen worden in linguïstische en culturele factoren, waarbij ze aantekenen dat deze factoren met elkaar kunnen samenhangen. Bij linguïstische bronnen van itembias spelen de taalelementen en de teksteigenschappen van de items en het bijbehorende contextmateriaal een centrale rol. Bij culturele bronnen van itembias gaat het vooral om verschillen in cultuurspecifieke voorkennis. Dit kan met name het geval zijn met de contexten die bij items gebruikt worden.

Het niet beheersen van wat een toets meet, kan evenwel ook een bron van itembias zijn. Verschillen tussen allochtone en autochtone leerlingen op het gebied van bijvoorbeeld rekenen kunnen itembias veroorzaken. Stel, dat om de een of andere reden het onderwijsaanbod voor allochtone leerlingen ten aanzien van de rekendomeinen Procenten en Verhoudingen beperkter of minder efficiënt is dan voor autochtone leerlingen en dat allochtone leerlingen meer profiteren van het onderwijs in de Hoofdbewerkingen (optellen, aftrekken, vermenigvuldigen en delen). Wanneer de leerlingen op basis van de rekentotaalscore naar niveau geclassificeerd worden, is het mogelijk dat een of meer items over Procenten en Verhoudingen partijdig zijn in het nadeel van

allochtone leerlingen en dat items over de Hoofdbewerkingen partijdig zijn in het voordeel van allochtone leerlingen. Deze bron van itembias vloeit voort uit de definitie van itembias. Op dit moment laten we in het midden of we hier te maken hebben met beperkingen ten aanzien van de constructvaliditeit van een toets. In hoofdstuk acht komen we hierop terug.

In het volgende overzicht worden drie oorzaken van itembias onderscheiden. In navolging van De Jong & Vallen (1989) onderscheiden we linguïstische (2.2.2) en culturele bronnen van itembias (2.2.3), toegevoegd worden bronnen die betrekking hebben op het gegeven onderwijs (2.2.4). Vooraf gaat een theoretisch raamwerk voor de relatie tussen taalvaardigheid en schoolsucces.

2.2.1 Een theoretisch raamwerk voor de relatie tussen taalvaardigheid en schoolsucces van allochtone leerlingen

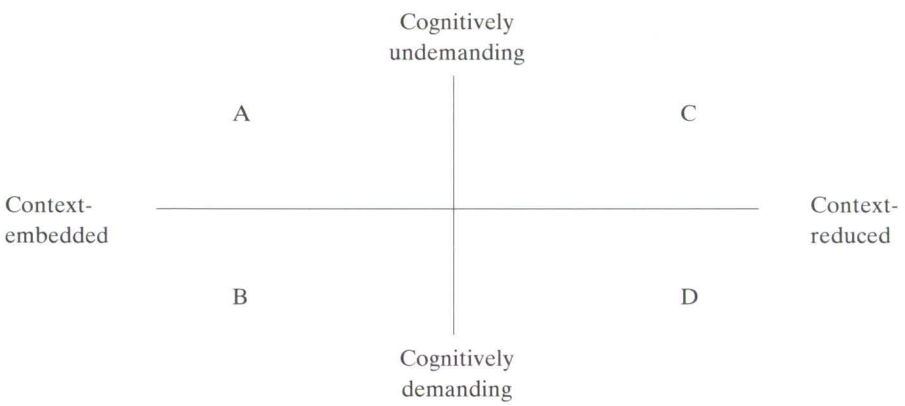
Cummins heeft een theoretisch raamwerk gepresenteerd dat in verband met het accent op de evaluatie van de vaardigheid in de eerste taal (T1) en de tweede taal (T2) belangwekkend is voor onderzoek naar itembias voor allochtone leerlingen (1979; 1984a; 1991a). Zijn gedachten over de afhankelijkheid van de ontwikkeling van T1 en T2 (Cummins, 1979) geven aan dat bij het zoeken naar de oorzaken van itembias de eerste taal niet buiten beschouwing mag blijven. Cummins' tweedimensionale model (1984a), met enerzijds aandacht voor de mate waarin het taalgebruik eisen stelt aan het cognitief functioneren en anderzijds voor de mate waarin het taalgebruik ondersteund wordt door een informatieve context, biedt wellicht mogelijkheden de dimensies in verband te brengen met mogelijke oorzaken van itembias.
We bespreken nu eerst Cummins' hypothese over de afhankelijkheid van de ontwikkeling van T1 en T2, daarna gaan wij in op zijn tweedimensionale model.

Volgens de '*afhankelijkheidshypothese*' is de vaardigheid in de eerst-verworven taal voor een groot deel ondersteunend voor de ontwikkeling van de vaardigheid in de tweede taal (Cummins, 1979; 1984a; 1991a; 1991b). Hiermee wordt deels een verklaring gegeven voor onderzoeksresultaten waarin T2-beheersing sterk blijkt samen te hangen met T1-beheersing (Verhoeven, 1987; Kerkhoff, 1988; Hacquebord, 1989; Diaz & Klinger, 1991). De samenhang tussen T1 en T2 wordt verklaard uit het niet gescheiden opgeslagen zijn van de cognitieve systemen van beide talen. Een deel van de taalvaardigheid in beide talen is universeel: de Common Underlying Proficiency (CUP). Bepaalde taalaspecten ontwikkelen zich universeel en vervullen een functie bij het verwerven van zowel T1 als T2. Dit zijn aspecten van taalvaardigheid die verbonden zijn met onderliggende, algemene cognitieve principes. Daarnaast onderscheidt Cummins meer taalspecifieke vaardigheden als uitspraak, 'fluency' en spelling. Er is volgens Cummins (1979) een zeker niveau van T1-beheersing nodig en een bepaald aanbod in T2 om een positieve transfer van T1 op T2 mogelijk te maken.

Er wordt vanuit verschillende invalshoeken kritiek op de afhankelijkheids-hypothese van Cummins geleverd (vgl. Genesee, 1984; Verhoeven, 1987; Kerkhoff, 1988; Hacquebord 1989). In de eerste plaats wordt er op gewezen dat correlationele verbanden niet zonder meer geïnterpreteerd kunnen worden als

causale relaties. Er kunnen wel correlaties tussen T1- en T2-vaardigheden gevonden worden, maar het is mogelijk dat een of meer andere onafhankelijke variabelen direct of indirect mede van invloed zijn op de T2-vaardigheid. Genesee (1984) stelt dat lage onderwijsresultaten van leerlingen uit lagere sociaal-economische milieus aanleiding geven te veronderstellen dat naast linguïstische ook sociaal-psychologische variabelen direct of indirect effect kunnen hebben op T2-verwerving. Daarnaast wordt ook gewezen op het feit dat de T2-beheersing invloed kan hebben op de T1-vaardigheid. Deze kritiek geeft in feite aan dat de effecten van T1 op T2 het beste samen met andere relevante variabelen in een causaal verklaringsmodel kunnen worden geschat. Hierdoor kan vastgesteld worden wat de effecten zijn van de onderscheiden variabelen op elkaar en op de T2-vaardigheid.

Het genoemde *tweedimensionale model* geeft Cummins (1984a: 12) als volgt weer.



De verticale dimensie geeft de mate aan waarin een taalgebruikstaak cognitief gezien hoge eisen stelt. De horizontale dimensie geeft de mate aan waarin dit taalgebruik is ingebed in een context waaraan de taalgebruiker ondersteunende informatie kan ontlenen.
Onder context verstaat Cummins de situationele informatie en feedback van andere taalgebruikers bij het verwerken van het taalaanbod. Het gaat om de mate van informatie die de taalgebruikssituatie begeleidt, extra betekenis aan het taalaanbod toevoegt en de T2-leerder houvast kan bieden bij het verwerken van het taalaanbod. Hierbij kan gedacht worden aan nonverbale stimuli en feedback-mechanismen die het mondeling taalgebruik kunnen begeleiden.
Bij context-arm taalgebruik moeten T2-leerders zich voornamelijk op de linguïstische betekenisdragers baseren. Zij moeten bij context-arm taalgebruik een beroep doen op hun kennis van de wereld om het taalaanbod op de juiste wijze te interpreteren. De enige context zijn de tekstelementen zelf.
Voorbeelden van communicatieve situaties, die van links naar recht op de horizontale as geplaatst worden, zijn: deelname aan een discussie, een brief schrijven aan een goede bekende, lezen of schrijven van een artikel in een vaktijdschrift. Door een beperkte T2-vaardigheid en een cultureel gebonden kennis van de wereld van allochtone leerlingen is voor hen context-arm taalaanbod in T2 vaak moeilijk.

Voor onderzoek naar itembias is deze dimensie belangrijk, omdat toetsitems over het algemeen in hoge mate 'context-reduced' zijn.
De verticale dimensie van Cummins' model heeft betrekking op de eerder door hem omschreven acroniemen 'CALP' en 'BICS'. CALP betekent Cognitive Academic Language Proficiency en BICS staat voor Basic Interpersonal Communicative Skills. Onder BICS worden meer informele taalvaardigheden verstaan als het kunnen spreken en luisteren in dagelijkse taalgebruikssituaties. BICS-vaardigheden hebben betrekking op concepten als uitspraak en 'fluency'. De CALP-vaardigheden zijn gekenmerkt door cognitief veeleisende meer formele taken, waarbij vooral gedacht moet vooral worden aan schriftelijke vaardigheden: lezen en schrijven. Wald (1984) stelt dat CALP-vaardigheden in feite de vaardigheden zijn waar toetsen voor het meten van T2-vaardigheden een beroep op doen. Hij stelt tegenover de CALP-vaardigheid de spontane taal in alledaagse mondelinge communicatieve situaties. Cummins (1984b) is van mening dat Walds interpretatie van CALP verwarrend is, omdat de toetssituatie slechts een operationalisatie van CALP-gedrag is. Maar Cummins is ook van mening dat een T2-toets in hoge mate 'cognitively demanding' en 'context-reduced' is.
Cummins (1984a; 1991a) wijst erop dat leerkrachten op basis van een gevorderde BICS-vaardigheid de neiging hebben om de taalvaardigheid van allochtone leerlingen in het algemeen relatief hoog in te schatten. BICS-vaardige leerlingen redden zich vrij goed in de mondelinge taalgebruikssituaties, spreken redelijk vloeiend T2. Cummins wijst erop dat een goede BICS-vaardigheid nog niet een goede CALP-vaardigheid hoeft te betekenen. De meer formele CALP-vaardigheid stelt hogere en andere eisen aan de taalvaardigheid. In het tweedimensionale model van Cummins moeten we de CALP-vaardigheden vooral rechtsonder (D) zoeken. De BICS-vaardigheden moeten in het linkerboven (A) deel geplaatst worden. Het verschil tussen A en D geeft het maximale onderscheid in het model aan. Volgens Cummins (1991a) beheerst een kind dat Engels als T2 leert, gemiddeld na 2 jaar taalcontact de BICS-vaardigheid op eenzelfde niveau als een leeftijdgenoot voor wie het Engels de moedertaal is. Gemiddeld heeft datzelfde kind 5 jaar of meer nodig om de CALP-vaardigheid op het niveau van zijn autochtone leeftijdgenoot te beheersen (Cummins, 1991a: 169).

Cummins benadrukt dat het model geen dichotome presentatie van de werkelijkheid pretendeert te zijn, maar juist glijdende schalen bevat. Toetsen zullen in het model vrijwel altijd bij D geplaatst moeten worden. In onderzoek naar itembias kan nagegaan worden in welke mate een item een beroep doet op CALP en in welke mate het item met name voor leerlingen uit etnische minderheidsgroepen context-arm is.

2.2.2 Potentiële linguïstische bronnen van itembias

De Jong & Vallen (1989: 392) stellen dat er slechts weinig bekend is over de vraag met welke specifieke T2-elementen allochtone leerlingen aan het einde van de basisschool problemen hebben. Wellicht zullen sommige gevorderde T2-leerders soortgelijke problemen hebben met de Nederlandse taal als autochtone leerlingen. Voor itembias is het evenwel van belang om te zoeken naar elementen van T2-verwerving waarbij de leerprestaties van allochtone en

autochtone leerlingen verschillen. Deze verschillen worden nu besproken op het niveau van woorden, zinnen en teksten. Als laatste potentiële linguïstisch bron van itembias komen metalinguïstisch vaardigheden aan de orde. Het zou immers niet uitgesloten kunnen zijn, dat allochtone leerlingen door hun tweetaligheid ten aanzien van bepaalde metalinguïstische vaardigheden in het voordeel zijn vergeleken met autochtone leerlingen.

a *Moeilijkheidsgraad van woorden*
De Nederlandse woordenschat lijkt voor T2-leerders een belangrijk struikelblok voor schoolsucces te vormen. Driessen (1990) geeft een overzicht van onderzoekers die geconstateerd hebben dat de kennis van de Nederlandse woordenschat van allochtone leerlingen achterblijft bij die van autochtone. Vermeer (1986) constateert dat de geschatte woordvoorraad van Turkse en Marokkaanse kinderen in de beginfase van het basisonderwijs beduidend afwijkt van die van autochtone leerlingen. Hij schat dat in die fase van het onderwijs de Nederlandse woordenschat van Turkse leerlingen bijna 1500 woorden kleiner is dan die van autochtone leerlingen. Verder merkt Vermeer op dat de relatieve afstand tussen allochtone en autochtone leerlingen met betrekking tot de produktieve woordenschat in de loop van de jaren niet kleiner wordt. De geschatte produktieve woordvoorraad neemt volgens hem gemiddeld elke week met 10 à 11 woorden toe, maar omdat deze groei bij allochtone en autochtone leerlingen in de onderzochte periode (6- tot 9-jarigen) ongeveer even groot is, halen de allochtone kinderen hun relatieve achterstand in die periode niet in (Vermeer, 1986). Verhoeven & Vermeer (1992) constateren dat de ontwikkeling van de receptieve woordenschat in de bovenbouw van het basisonderwijs sneller verloopt dan in de onderbouw. Zij verklaren dit uit de toenemende leesvaardigheid van de leerlingen. Ook Verhoeven en Vermeer geven aan dat de verschillen tussen de omvang van de woordenschat van allochtone en autochtone leerlingen naar verhouding in de loop van het basisonderwijs ongeveer constant blijven.
Uit het onderzoek van De Jong (1987) blijkt dat de Nederlandse woordenschat van allochtone leerlingen in Rotterdam en omgeving zowel aan het einde van het basisonderwijs als in de beginfase van het voortgezet onderwijs duidelijk achterblijft bij die van autochtone leerlingen. Hij geeft tevens aan dat de receptieve woordenschat van de leerlingen die naar MAVO, HAVO en VWO gaan, in de beginfase van het voortgezet onderwijs sterker toeneemt dan die van de IBO- en LBO-leerlingen. Dit geldt voor autochtone maar vooral voor allochtone leerlingen. Hacquebord (1989) vindt dat de woordenschat van Turkse LBO- en MAVO-leerlingen kleiner is dan die van Nederlandse leerlingen uit dezelfde onderwijstypen. Zij constateert dat de woordenschat van de Turkse leerlingen die naar het MAVO gaan, sterker groeit dan die van de Nederlandse klasgenoten. Net als De Jong treft Hacquebord de laagste groeiscores aan bij de LBO-ers van zowel allochtone als autochtone herkomst. Uit het bovenstaande kan geconcludeerd worden dat de kans bestaat dat aan het einde van het basisonderwijs de receptieve en produktieve woordenschat van allochtone en autochtone leerlingen verschilt. Dit is een belangrijk gegeven in verband met mogelijke bronnen van itembias.

Een lage *woordfrequentie* kan van invloed zijn op de moeilijkheidsgraad van een woord. Frequente woorden worden sneller en beter geleerd en herkend dan

infrequente woorden (Taylor & Taylor, 1990). Volgens De Jong & Vallen (1989) bestaat er echter niet altijd een rechtstreeks verband tussen de lagere frequentie van woorden en het kunnen aangeven van de betekenis van een woord. Sommige hoogfrequente woorden kunnen moeilijk zijn, omdat hun betekenis varieert naargelang de context waarin ze voorkomen. Verhoeven & Vermeer (1992) wijzen erop dat de informatiewaarde van hoogfrequente woorden, zoals bepaalde functiewoorden, niet altijd hoog is. Het lidwoord 'de' is bijvoorbeeld hoogfrequent, maar de informatie die van het woord op zich (zonder context) uitgaat is relatief gering.

Uit het voorgaande is gebleken dat de *context* waarin een bepaald woord wordt gebruikt van groot belang is voor de moeilijkheidsgraad ervan. De context kan verschillen in de mate waarin deze aan leerlingen verschillende woorden ontlokt, bijvoorbeeld wanneer leerlingen worden gevraagd om zinnen af te maken of om in zinnen ontbrekende woorden in te vullen. Zo kunnen er in de zin 'In het dal zag hij ....' meer woorden ingevuld worden dan in de zin 'Jan bewaart zijn kleren in de ....'. Volgens Schwanenflugel (1991) blijkt uit empirisch onderzoek dat wanneer een woord nauw gerelateerd is aan de context de moeilijkheidsgraad van het woord over het algemeen lager is. Het is denkbaar dat allochtone leerlingen minder steun aan de context hebben om de juiste betekenis van een woord te kiezen dan autochtone leerlingen.
Voor de moeilijkheidsgraad van een woord is het ook van belang waar het woord geplaatst moet worden op de *dimensie abstract-concreet*. Concrete woorden verwijzen meer naar voorstelbare verschijnselen en worden door leerlingen beter gedefinieerd en ze herinneren zich deze woorden beter dan abstracte woorden (Taylor & Taylor, 1990).
Ambigue woorden kunnen ook oorzaak van itembias zijn. Ambigue woorden zijn volgens Rayner & Morris (1991) woorden waarbij meer dan één betekenis van het woord geactiveerd wordt. Ambigue woorden komen in feite zeer veel voor en de context verleent het woord doorgaans zijn specifieke betekenis. Het is vaak moeilijker om de betekenis van een geïsoleerd woord aan te geven, dan de betekenis van een ambigu woord in een context (Tabossi, 1991). Rayner & Morris (1991) vonden dat lezers in eerste instantie één betekenis van het woord kiezen, nadat mogelijke betekenissen zijn verkend. Wanneer de oorspronkelijk gekozen betekenis niet de juiste lijkt te zijn, analyseert de lezer de zin opnieuw of zoekt hij verderop in de tekst naar aanknopingspunten (Rayner & Morris, 1991; Swinney, 1991). Het is moeilijk om te achterhalen welk deel van de zin, nadat een woord is herkend, bepaalt welke betekenis aan een woord wordt toegekend (Tobassi, 1991). Gernsbacher & Faust (1991) geven aan dat het voor de keuze van de juiste betekenis van een woord van belang is dat de lezer in staat is om niet passende betekenissen van dat woord te onderdrukken.
Uit het bovenstaande blijkt dat de context een grote invloed heeft op de moeilijkheidsgraad van woorden. Dit gegeven is voor onderzoek naar potentiële bronnen van itembias belangrijk.

b *Moeilijkheidsgraad van zinnen*
Het doorgronden van de syntactische structuur van zinnen vormt een voorwaarde voor het ordenen, interpreteren en reconstrueren van de semantische inhoud ervan. Inzake het begrijpen van zinsstructuren spelen verschillende typen conventies een belangrijke rol. In verband met itembias

moeten we er rekening mee houden dat leerlingen van Nederlandse herkomst vertrouwder met deze conventies zijn dan allochtone leerlingen (vgl. De Jong, 1987; Hacquebord, 1989). Kerkhoff (1988) heeft in haar onderzoek naar de T2-vaardigheid van allochtone en autochtone leerlingen aan het einde van de basisschool opstellen van beide groepen leerlingen geanalyseerd. Zij vond dat Turkse, Marokkaanse, Molukse en Surinaamse leerlingen meer syntactische fouten maakten dan autochtone leerlingen. Het niet gebruiken van één of meer noodzakelijk vereiste woorden in een zin (bijvoorbeeld: "Charlie die begreep dat _ zijn kans was") en het gebruiken van verkeerde voegwoorden (bijvoorbeeld: "Charlie kijkt een beetje sip en (= want) hij is bijna blut") zijn in haar onderzoek de meest voorkomende syntactische fouten. O'Malley & Chamot (1990) stellen dat competente taalgebruikers vooral gericht zijn op de betekenis van een zin en dat T2-leerders aandacht moeten wijden aan zowel de betekenis als de structuur van de zin. Moedertaalsprekers merken bij veranderingen in zinnen meer de aanpassingen in de betekenis van de zin op, terwijl niet-moedertaalsprekers beter kunnen aangeven wanneer de vorm van de zin is gewijzigd.

Taylor & Taylor (1990) geven indicaties voor de vraag welke soorten zinnen relatief moeilijk zijn te begrijpen en/of te produceren voor allochtone leerlingen in vergelijking met autochtone.
Ontkennende zinnen zijn moeilijker dan bevestigende. Leerlingen hebben meer tijd nodig om op ontkennende zinnen te reageren en maken hierbij meer fouten. Dit geldt in het bijzonder voor zinnen met een dubbele ontkenning.
Passieve zinnen zijn voor leerlingen moeilijker te begrijpen en te produceren dan actieve zinnen. Actieve zinnen zijn korter en syntactisch minder complex dan passieve. Van belang is ook welk type werkwoord in de passieve zin gebruikt wordt. Leerlingen begrijpen passieve zinnen met werkwoorden, die naar een statische situatie verwijzen (bijvoorbeeld: vasthouden) minder gemakkelijk dan passieve zinnen met een werkwoord dat naar een actie verwijst (bijvoorbeeld: slaan).
Figuurlijk taalgebruik is voor kinderen over het algemeen moeilijk, omdat ze zich vooral op de letterlijke betekenis richten. Naarmate ze ouder worden, neemt de toegankelijkheid voor de betekenis van metaforen, gezegden en specifieke idiomatisch taalgebruik toe (Taylor & Taylor, 1990; Johnson, 1991). Het is denkbaar dat de toename van kennis en inzicht in figuurlijk taalgebruik voor een belangrijk deel verklaard kan worden door de bijdrage die het onderwijs hieraan levert. Johnson (1991) concludeert op basis van haar onderzoek dat verwacht mag worden dat T2-leerders niet in het voordeel maar ook niet in het nadeel zijn bij het interpreteren van figuurlijk taalgebruik vergeleken met monolinguale leerlingen. Volgens Cacciari & Glucksberg (1991) maakt het wel verschil in welke mate er afstand bestaat tussen de figuurlijke betekenis van een uitdrukking en de letterlijke. Zij wijzen erop dat de moeilijkheidsgraad van figuurlijk taalgebruik ook afhankelijk is van de mate waarin de context de interpretatie van de uitdrukking ondersteunt. Ook hier wordt gewezen op het belang van de context voor het aangeven van de juiste betekenis. Opgemerkt moet worden dat allochtone leerlingen bij figuurlijk taalgebruik over het algemeen minder steun aan de context zullen hebben dan autochtone leerlingen, hetgeen een bron van itembias zou kunnen zijn.

c *Moeilijkheidsgraad van teksten*
Voor leerlingen uit etnische minderheidsgroepen lijkt het van belang om bij de moeilijkheidsgraad van teksten onderscheid te maken naar het niveau van de teksten. Hacquebord (1989) maakt onderscheid tussen het micro- (woord- en zinsniveau), meso- (alinea-niveau) en macroniveau (het hoofdthema, de tekstsoort, de strekking) van een tekst. Uit haar onderzoek blijkt dat 10% van de variantie in tekstbegripscores op microniveau door de etnische achtergrond wordt verklaard, op mesoniveau is dat 2% en op macroniveau 1%. Dat op het microniveau de meeste variantie door etniciteit wordt verklaard, sluit aan bij de relatief grote en algemeen aangetoonde verschillen in T2-woordkennis. De geringe verschillen op macroniveau worden door Hacquebord in verband gebracht met efficiënte leesstrategieën die Turkse leerlingen hanteren om hun geringere T2-kennis op microniveau te compenseren.

Verhoeven & Vermeer (1992) wijzen eveneens op de functie van de woordenschat bij het lezen van teksten. Zij merken op dat autochtone leerlingen onbekende woorden in een tekst raden met behulp van de woorden er omheen, waarvan ze de meeste kennen. Voor allochtone leerlingen zitten er vaak zoveel onbekende woorden ('gaten') in een tekst dat zij de onbekende woorden niet of nauwelijks kunnen afleiden uit de tekst. Verhoeven & Vermeer (1992) stellen dat voor allochtone leerlingen in teksten niet meer dan 10% van de woorden onbekend mag zijn.
Nienhuis (1991) ging na hoeveel procent van de woorden in een tekst onbekend mag zijn om een tekst te kunnen lezen en begrijpen. Onbekende woorden zijn in dit onderzoek de ontbrekende woorden in een tekst. In Engelse en Franse teksten werden de 10%, respectievelijk 25% minst frequente woorden weggelaten ('gatentekst') alvorens VWO-5 leerlingen vragen over de hoofdgedachte van (delen van) de tekst moesten beantwoorden. Nienhuis concludeert dat bij teksten waarin 75% van de meest frequente woorden blijven staan (een dekkingspercentage van 75%), er nauwelijks sprake is van globaal begrip van de tekst, terwijl bij teksten waarin 90% van de meest frequente woorden zijn overgebleven globaal tekstbegrip maar zeer beperkt het geval is. Het onderzoek van Nienhuis maakt eveneens duidelijk dat het microniveau van een tekst van groot belang kan zijn voor tekstbegrip. Opgemerkt moet worden dat 'onbekend' en 'ontbrekend' ten aanzien van tekstbegrip differentiële effecten kan hebben en dat het leren van een moderne vreemde taal door VWO-leerlingen niet hetzelfde is als het leren van Nederlands als T2 door allochtone leerlingen in het basisonderwijs.

Whitney & Waring (1991) merken op dat het *geheugen* een rol kan spelen bij het begrijpen van teksten. Leerlingen met een goed geheugen slaan verschillende interpretaties van een bepaald stuk tekst op in het geheugen. Vervolgens gaan deze leerlingen na welke interpretatie geldig is, wanneer in de volgende delen van de tekst nieuwe informatie wordt gegeven. Leerlingen met een minder goed functionerend geheugen interpreteren elke gebeurtenis in de tekst meer geïsoleerd van de rest van de tekst en bovendien hebben zij zich relatief vroeg in de tekst een idee gevormd over de hoofdgedachte van de tekst.

Taylor & Taylor (1990) stellen dat tekstbegrip ook afhankelijk is van de mate waarin de betekenis van de tekst *plausibel* is. Zij onderscheiden plausibele,

niet-plausibele en neutrale tekstfragmenten. Tekstfragmenten worden bij deze indeling geclassificeerd naar de mate waarin de gebeurtenissen in de tekst aannemelijk, voorstelbaar en geloofwaardig zijn.

Op het niveau van de teksten zijn verder de volgende twee potentiële bronnen van itembias belangrijk: referenties in teksten en tekstsignalen.
Referenties in teksten, verwijzingen naar elementen binnen dezelfde zin of in voorafgaande of volgende zinnen, kunnen aanzienlijk verschillen in moeilijkheidsgraad. Vooral minder taalvaardige leerlingen blijken moeite te hebben met verwijswoorden. Zij hebben meer problemen met het onderdrukken van irrelevante informatie in de tekst bij het leggen van relaties tussen verwijswoord en referent. (Gernsbacher & Faust, 1991). Moeilijke referenties zijn volgens De Jong & Vallen (1989: 395)
- verwijzingen waarbij als bekend vooronderstelde informatie (Given) door de schrijver niet voor de nieuw bedoelde informatie (New) geplaatst wordt;
- verwijzingen waarbij over relatief grote tekstdelen heen wordt verwezen (vgl. Taylor & Taylor, 1990);
- ambigue verwijzingen (vgl. Taylor & Taylor, 1990);
- referenties waarbij het verwijswoord getalsmatig afwijkt van de referent (bijvoorbeeld: verwijswoord = ze; referent = het gezin);
- clausale (verwijzingen naar een zin of naar elementen uit een zin) en werkwoordelijke (verwijzingen naar een werkwoord meestal in een andere zin) referenties.

Tekstsignalen zijn middelen om de structuur van een tekst te verhelderen (bijvoorbeeld: in het volgende hoofdstuk worden de resultaten gegeven). Teksten zonder expliciete opmerkingen over de structuur van de tekst kunnen problemen opleveren voor minder taalvaardige leerlingen.
De Jong & Vallen (1989) constateren dat in een tekst uit de Eindtoets Basisonderwijs 1987 een niet-gangbare structuuraanduider voorkomt en zij merken op dat een titel boven een tekst uit dezelfde toets niet past bij de tekstinhoud. De Jong & Vallen (1989: 396) stellen dat het voorstelbaar is dat ongebruikelijke of onjuiste structuuraanduiders itembias kunnen veroorzaken voor allochtone leerlingen.

d *metalinguïstische vaardigheden*
Het feit dat leerlingen uit etnische minderheidsgroepen tweetalig zijn, kan in vergelijking met monolinguale leerlingen een positief effect hebben op het leren van bepaalde T2-elementen. Uit onderzoek blijkt dat de metalinguïstische vaardigheid en het metalinguïstisch bewustzijn van tweetalige leerlingen zich over het algemeen beter ontwikkelen dan die van monolinguale (Hakuta, 1986; Taylor & Taylor, 1990; Cummins, 1991a; 1991b; Bialystok, 1991). Verschillen tussen tweetalige en monolinguale leerlingen op het terrein van metalinguïstische vaardigheden zijn in verband met onderzoek naar itembias van belang, omdat deze verschillen potentiële bronnen van itembias zijn.

De definities van metalinguïstische vaardigheid zijn over het algemeen abstract van aard waardoor het niet altijd precies duidelijk wordt in welk aspect van taalvaardigheid T2-leerders beter zijn dan monolinguale leerlingen. Eerst gaan we in op enkele definities van metalinguïstische vaardigheden.

Taylor & Taylor (1990: 268) geven aan dat het bij metalinguïstische vaardigheid gaat om de bekwaamheid om te denken en te spreken over de taal als object. Volgens Cummins (1991a: 164) gaat het bij deze vaardigheid om de expliciete kennis over de structuur en functie van de taal zelf. Diaz & Klingler (1991) stellen dat het metalinguïstisch bewustzijn betrekking heeft op de taal als een objectief en arbitrair tekensysteem. Bialystok (1991: 113) stelt dat onderzoekers met het begrip metalinguïstisch bewustzijn doelen op het ingevoerd zijn in de regels van de taal met haar eigen eigenschappen en structuur. Bialystok beschrijft de metalinguïstische ontwikkeling aan de hand van twee componenten. Elke component is verantwoordelijk voor een aspect van een taalhandeling en voor de ontwikkeling daarvan. De eerste component is verantwoordelijk voor de wijze waarop het kind de taal mentaal representeert. De mentale representatie heeft betrekking op het gebruik van symbolen die verwijzen naar een bepaalde betekenis. De leerling reflecteert op een doorgaans automatisch verlopende taalhandeling die op dat moment echter als een object wordt beschouwd (vgl. Sijtstra, 1991). Bij Bialystok (1991: 117) is deze reflectie vooral gericht op de relatie tussen symbolen en woorden of concepten. De mentale taalrepresentatie van kinderen wordt meer en meer gestructureerd en geëxpliciteerd. De eerste component wordt door Bialystok (1991) 'analyse van linguïstische kennis' genoemd. De tweede component is de aandacht-functie, die de informatie van de mentale representatie selecteert met het doel die in een bepaalde context te gebruiken. De selectieve aandacht groeit naar mate de linguïstische kennis van de leerling groter wordt. De tweede component, 'controle van de linguïstische handeling' heeft vooral betrekking op het vermogen om de aandacht ergens op te richten, maar veronderstelt wel een bepaalde basis aan linguïstische kennis. Zo is voor het lezen een goede balans tussen beide componenten nodig: de aandacht die gericht moet zijn op zowel de woorden als op de betekenis ervan. Wanneer leerlingen toetsitems beantwoorden wordt in grote mate een beroep gedaan op de beide componenten van metalinguïstische ontwikkeling.
In de literatuur worden een aantal metalinguïstische (deel)vaardigheden genoemd die allochtone leerlingen beter zouden beheersen dan autochtone.

Taylor & Taylor (1990) geven aan dat T2-leerders over het algemeen een beter ontwikkeld metalinguïstisch bewustzijn hebben. T2-leerders blijken tijdens het aanvankelijk lezen beter dan monolinguale kinderen inzicht te hebben in de samenstelling van woorden. Zij kunnen in een tekst beter tweelettergrepige en meerlettergrepige woorden identificeren, terwijl tweetalige en monolinguale leerlingen even goed zijn in het aanwijzen van éénlettergrepige woorden. Taylor & Taylor (1990: 343) zeggen verder dat goed uitgevoerd T2-onderwijs geen nadelig effect heeft op de cognitieve en linguïstische vaardigheid. T2-leerders kweken een grotere linguïstische en cognitieve flexibiliteit en veelzijdigheid.

Bialystok (1991) noemt ook twee punten waarbij T2-leerders in het voordeel zijn op monolinguale leerlingen.
- Door ervaring opgedaan met lezen en schrijven in T1 en T2 hebben T2-leerlingen een voorsprong op monolinguale leerlingen bij metalinguïstische taken die in sterke mate een beroep doen de analyse van kennis. Hierbij gaat het bijvoorbeeld om taken waarbij leerlingen grammaticale onjuistheden in zinnen moeten opsporen. Het gaat hier dus om

het opsporen en niet om het corrigeren van grammaticaal onjuist gevormde zinnen. Er zijn geen aanwijzingen dat T2-leerders beter zijn dan monolinguale kinderen in het verbeteren van onjuist gevormde zinnen (Johnson, 1991).
- Wanneer leerlingen ervaring hebben opgedaan met spreken en luisteren in T1 en T2, dan hebben T2-leerlingen een voorsprong op monolinguale leerlingen bij metalinguïstische taken die in sterke mate aandacht voor en controle op het taalgebruik als zodanig vereisen. Hierbij kan het om taken gaan waarbij leerlingen grammaticale onjuistheden moeten opsporen in zinnen die grammaticaal correct zijn, maar inhoudelijk onregelmatigheden bevatten (bijvoorbeeld: Waarom blaft de kat zo luid?). Dit onderzoek wijst erop dat T2-leerders de betekenis en de vorm van T2 beter kunnen scheiden.

Bialystok (1991) geeft verder aan dat T2-leerlingen beter zijn in het geven van formele definities (bijvoorbeeld: Wat is een woord? Hoe kun je een woord herkennen?).

Volgens Diaz & Klingler (1991) blijkt uit empirisch onderzoek dat T2-leerders een aantal metalinguïstische vaardigheden beter beheersen dan monolinguale kinderen. Zij noemen
- het opsporen van ambigue woorden en tautologieën;
- het gevoel voor syntactische juiste zinsstructuren;
- het kunnen aangeven waar twee talen door elkaar gebruikt worden.

Johnson (1991) stelt dat T2-leerders een voorsprong hebben op monolinguale kinderen bij het maken van non-verbale taken waarbij de oorspronkelijke organisatie van het stimulusmateriaal misleidend is en een reorganisatie van de stimuli noodzakelijk is voor het uitvoeren van de taak. Deze taken vereisen een wendbaarheid in het aanpassen van de gegeven structuur. Volgens Johnson wordt deze algemene cognitieve flexibiliteit vooral veroorzaakt door de context waarin T2 geleerd en gebruikt wordt en waarbij het wisselen tussen taalsystemen noodzakelijk is.

Volgens Diaz & Klingler zijn T2-leerders eveneens beter in classificeren, creativiteit, analogie-redeneringen en in visuele-ruimtelijke vaardigheden dan monolinguale leerlingen. Het is volgens hen nog niet duidelijk hoe deze cognitieve vaardigheden gerelateerd zijn aan metalinguïstische vaardigheden (Diaz & Klingler, 1991).

Cummins (1991a) concludeert dat het verwerven van T2 geen negatieve invloed heeft op de linguïstische en intellectuele ontwikkeling. Hoewel er geen afdoende empirisch fundament voor is, zijn er aanwijzingen dat er subtiele voordelen zijn op metalinguïstisch en intellectueel terrein. Diaz & Klingler (1991) zijn het in deze met Cummins eens, maar ze tekenen hierbij aan dat de positieve effecten van het leren van T2 op de cognitieve ontwikkeling zich openbaren bij de vroege stadia van T2-ontwikkeling om later weer te verdwijnen. Deze gegevens zijn ontleend aan situaties waarin T1 en T2 simultaan aangeleerd worden en als evenwaardig beschouwd kunnen worden (additive situations). Er zijn volgens Diaz & Klingler (1991) aanwijzingen dat T2-verwerving negatieve effecten op de cognitieve ontwikkeling heeft in situaties waarbij het leren van T2 gepaard gaat met verlies van T1 (subtractive situations). In het proces waarin T1 geleidelijk vervangen wordt door T2 kan er

een periode optreden waarin de leerling in feite beide talen gebrekkig beheerst. Het is echter niet duidelijk of de negatieve effecten op de cognitieve ontwikkeling een permanent karakter hebben.

2.2.3 Potentiële culturele bronnen van itembias

De Jong & Vallen (1989) onderscheiden twee bronnen van culturele itembias: de culturele lading van teksten en de cultureel bepaalde toetservaring. Bij de culturele lading van teksten gaat het om de voorkennis van het contextmateriaal dat gebruikt is voor de toetsitems. In verband met de vele soorten contexten, die in toetsen gebruikt worden, lijkt het beter om van 'voorkennis van het contextmateriaal' te spreken.

a *voorkennis van het contextmateriaal*
Uit onderzoek naar de moeilijkheidsgraad van tekstbegriptoetsen blijkt dat verschillen tussen de scores van allochtone en autochtone leerlingen voor een deel verklaard kunnen worden door verschillen in *voorkennis* van de onderwerpen die in de tekst aan de orde worden gesteld. Maureau (1979) geeft aan dat zinnen met vergelijkbare syntactische structuren door de daarbij vereiste kennis van de wereld toch kunnen verschillen in begrijpelijkheid. Whitney & Waring (1991) gaan er van uit dat het begrijpen van een tekst afhankelijk is van de interactie tussen de stimuli die een tekst bevat en de sterkte van hetgeen reeds in het geheugen van een persoon aanwezig is. Het tekstbegrip is een functie van de stimuli van een tekst en de reeds aanwezige kennis en inzichten (vgl. O'Malley & Chamot, 1990; Bügel & Glas, 1991). Whitney & Waring (1991) en Bügel (1991) merken op dat teksten verschillen in de mate waarin specifieke voorkennis wordt verondersteld. Tekstbegriptoetsen doen een beroep op kennis van de wereld die verondersteld wordt bij vrijwel iedere persoon uit de doelgroep aanwezig te zijn. Kuhlemeier & Van den Berg (1991: 144) wijzen erop dat bij het meten van taalvaardigheid rekening gehouden moet worden met taakeffecten en opdrachtspecificiteiten. Luisteren en lezen zullen bijvoorbeeld hoger correleren wanneer dezelfde tekst voor de metingen gebruikt wordt, dan wanneer luisteren en lezen met verschillende teksten getoetst worden. Kuhlemeier & Van den Berg (1991) doelen hier niet alleen op de voorkennis die een tekst veronderstelt, maar ook op het effect van het teksttype, het vraagtype en de taaksituatie. Zij adviseren deze aspecten op systematische wijze in een evaluatie-instrument te variëren. Zij melden dat de opdrachtspecificiteit het kleinst is bij leesopdrachten die conventioneel tekstbegrip meten en waarbij uitsluitend gebruik gemaakt wordt van meerkeuzevragen. De gemiddelde ware variantie in opdrachtspecificiteit bestaat bij dit soort toetsen voor 80% uit variantie in taalvaardigheid. Zij verklaren dit relatief hoge percentage verklaarde variantie uit het effect van het onderwijsaanbod, waardoor vrijwel alle leerlingen bekend zijn met het fenomeen 'tekst plus meerkeuzevragen'. Het percentage onverklaarde variantie (20%) schrijven zij toe aan verschillen in voorkennis van of affiniteit met het onderwerp dat in de tekst aan de orde wordt gesteld.
Hacquebord (1989) heeft in haar onderzoek toetsen ontwikkeld, die geacht worden de benodigde voorkennis van een tekst te meten. Zij stelt echter, dat het zeer moeilijk is om de voorkennis te meten van alle onderwerpen die in teksten van tekstbegriptoetsen voorkomen. Bovendien is het niet eenvoudig om

nauwkeurig vast te stellen welke voorkennis een bepaalde tekst veronderstelt en op welke wijze deze voorkennis het beste in items geoperationaliseerd kan worden. Het ligt voor de hand dat in een voorkennistoets een scala van onderwerpen voorkomt, hetgeen meestal afbreuk doet aan de psychometrische kwaliteit van de toets. Hacquebord vindt de interne betrouwbaarheid van de door haar ontwikkelde voorkennistoets aan de lage kant (29 driekeuze-items; n=309; Cronbachs alpha .68).

Kerkhoff & Vallen (1985) vonden eveneens aanwijzingen voor de invloed van voorkennis. Opgemerkt moet worden dat Hacquebord voorkennis opvat als de algemene kennis die elke persoon met enige schoolopleiding van een bepaald onderwerp in de tekst heeft, terwijl Kerkhoff & Vallen vooral doelen op de herkenbaarheid van de tekst in verband met de culturele achtergrond van de leerling: de culturele lading van teksten. Zij vergeleken de taalvaardigheid-scores van Turkse, Molukse en Nederlandse leerlingen op Nederlandstalige toetsen met een vergelijkbare moeilijkheidsgraad maar met een verschillende culturele lading. Zij constateerden dat elke etnische groep de beste resultaten behaalde op de toets die qua inhoud het meest correspondeerde met haar culturele achtergrond. Uit onderzoek van Johnston (1984) blijkt dat de voorkennis ook effect kan hebben op de tekstbegripscores van stads- versus plattelandskinderen.
Bügel & Robben-Willems (1989) en Bügel & Glas (1991) hebben in hun onderzoek naar itembias in de Centraal Schriftelijke Eindexamens moderne vreemde talen items gezocht die partijdig zijn voor jongens en meisjes. Zij vonden dat jongens door items bevoordeeld worden die betrekking hebben op tekstonderdelen die handelen over
- techniek;
- apparaten;
- gemotoriseerde vervoermiddelen;
- misdaad;
- sport;
- politieke en economische onderwerpen.

Meisjes maken items beter over
- intermenselijke relaties;
- gezinsproblemen;
- gevoelens.

De onderzoeken van Bügel & Robben-Willems en Bügel & Glas betreffen items die de vaardigheid tekstbegrip in het Frans, Duits of Engels meten. Eerst sporen zij met een statistische procedure partijdige items op. Vervolgens analyseren zij de inhoud van de tekst of het tekstgedeelte waar het item betrekking op heeft. Onderwerpen in teksten of tekstgedeelten zijn partijdig wanneer ze bij partijdige items horen. Bügel & Glas stellen dat er geen verband bestaat tussen vraagsoort en itembias.

De Jong & Vallen (1989) en Hacquebord (1989) wijzen op de problemen die zich voordoen bij het vaststellen van de voorkennis die voor een tekst nodig is en bij het operationaliseren van deze voorkennis in een meetinstrument. Het probleem wordt nog complexer omdat het in toetsen doorgaans niet om één tekst, maar om een aantal teksten gaat. Bovendien worden niet alleen teksten als contextmateriaal gebruikt, maar ook aardrijkskundige kaarten, tabellen,

grafieken, figuren en symbolen. Contextmateriaal doet dus ook een beroep op vaardigheden die in het onderwijs minder expliciet onderwezen worden, zoals ruimtelijk voorstellingsvermogen, schematische presentaties van de werkelijkheid en abstractievermogen. Dit contextmateriaal doet een beroep op een vaardigheid die Cummins (1984a) 'cognitively demanding' en 'context-reduced' noemt en dus volgens zijn theorie in het algemeen moeilijk is voor allochtone leerlingen. Het is onvermijdelijk dat de ene leerling meer voorkennis of zelfs meer emotionele binding heeft met het ene contextelement dan de andere leerling. Deze verschillende effecten kunnen tegen elkaar wegvallen, maar kunnen ook cumuleren waardoor de kans groter wordt dat het context-materiaal onbedoeld de toetsscores beïnvloedt. Dit kan met name het geval zijn bij leerlingen waarvan de culturele achtergrond sterk verschilt van die van de dominante groep leerlingen.

b *cultureel bepaalde toetservaring*

Cultureel bepaalde toetservaring is de tweede potentiële culturele bron van itembias. Het gaat hierbij om de ervaring in het maken van toetsen inclusief de kennis van de soort taken die in dergelijke toetsen te verwachten zijn en kennis van effectieve oplossingsstrategieën. Het gaat hierbij niet om de te meten vaardigheid en het contextmateriaal als zodanig, maar om de vertrouwdheid met het stimulusmateriaal en met de aard van de gewenste responsen. Scheuneman (1988) zegt dat cultureel bepaalde toetservaring als oorzaak van itembias drie achtergronden kan hebben: verschillen tussen allochtone en autochtone leerlingen met betrekking tot metacognitieve processen, persoonlijkheidsvariabelen en 'testwiseness'.

- Als voorbeelden van metacognitieve processen noemt Scheuneman 'vaststellen welk probleem opgelost moet worden, selecteren van voor de oplossing van het probleem belangrijke en onbelangrijke elementen, selecteren van een oplossingsstrategie, oplossingen afwegen en uitvoeren, gekozen oplossingen evalueren'. Vooral bij het uitvoeren van nieuwe taken kunnen daarbij verschillen tussen leerlingen zichtbaar worden.
- Met persoonlijkheidsvariabelen doelt Scheuneman op niet-cognitieve persoonskenmerken als temperament, angsten, motivatie, affectie en persoonlijke stijl. Deze persoonskenmerken kunnen het cognitief functioneren zowel positief als negatief beïnvloeden, waardoor ze aanleiding kunnen vormen voor itembias.
- Met 'testwiseness' doelt Scheuneman op het functioneren in de toetssituatie zelf. Vertrouwdheid met het op formele wijze omgaan met schriftelijk toetsmateriaal kan bij verschillende etnische groepen in ongelijke mate aanwezig zijn (vgl. Van de Vijver, Willemse & Van de Rijt, 1993).

Bij testen en toetsen is het gebruikelijk om, voordat de feitelijke afnamesituatie begint, de leerlingen een aantal opgaven ter kennismaking te laten maken. Door alle toetsdeelnemers ervaring op te laten doen met de wijze waarop de toets een appèl doet op hun vaardigheid, wordt verwacht dat de onbekendheid met de taak geen invloed heeft op de uiteindelijke toetsscores. Uit onderzoek is weinig bekend over de hoeveelheid ervaring die verschillende groepen vooraf moeten opdoen.

Er wordt wel onderzoek gedaan naar de effecten van itemtypen en van de volgorde van items in toetsen. Onderzoeken naar itemtype (open/gesloten items, twee/vierkeuze items) worden door Scheuneman & Steinhaus (1987) bekritiseerd, omdat de moeilijkheidsgraad van verschillende itemtypen moeilijk vergeleken kan worden en omdat niet zeker is of de items op dezelfde wijze het construct meten. Volgens Scheuneman & Steinhaus (1987) blijkt het veranderen van de volgorde van items doorgaans geen effect te hebben op de moeilijkheidsgraad van het item. Wanneer er wel effecten worden gevonden, dan betreft het meestal items aan het einde van de toets. Deze items kunnen in verband met tijdgebrek of vermoeidheid fout gemaakt zijn.
Uiteraard zal het moeilijk zijn om a posteriori de volgorde als oorzaak van itembias aan te wijzen. Om zicht te krijgen op de effecten hiervan is experimenteel onderzoek nodig.

2.2.4 Onderwijsaanbod als potentiële bron van itembias

Items kunnen tenslotte ook nog partijdig zijn omdat ze betrekking hebben op onderwijsdoelen die niet voor alle doelgroepen geldig zijn. Uit onderzoek blijkt dat leerkrachten niet aan alle leerlingen dezelfde eisen stellen als het gaat om het beheersen van leerstof. Jungbluth (1985) stelt dat leerkrachten nauwkeurig kunnen aangeven op welk prestatieniveau er in het onderwijsaanbod wordt gemikt. Bij leerlingen met een VWO-perspectief wordt een ruim onderwijsaanbod gerealiseerd en worden relatief hoge eisen gesteld. Deze leerlingen krijgen uitgebreid onderwijs in woordbenoeming, zinsontleding, het hanteren van tabellen en grafieken en er worden hoge eisen gesteld aan correcte spelling, geheugenwerk en historisch besef. Voor leerlingen met een LBO-perspectief zijn de eisen bescheidener en er wordt volstaan met leerstof die nodig wordt geacht voor later. Jungbluth wijst erop dat differentiatie met betrekking tot onderwijsdoelen vergaande gevolgen kan hebben voor de schoolloopbaan van leerlingen. Het advies voor het voortgezet onderwijs is volgens hem gedeeltelijk een uitvloeisel van het feit dat bepaalde leerlingen reeds lang onderwijs ontvingen dat afgestemd was op die vervolgkeuze. Leerkrachten kwalificeren volgens Jungbluth (1985: 134) leerlingen die het HAVO/VWO-niveau niet halen als leerlingen voor wie een andere succesformule geldt. Van der Hoeven-van Doornum (1990) vond dat leerlingen aan wie lagere eisen worden gesteld, vooral afkomstig zijn uit gezinnen met een (volgens de leerkracht) gering onderwijsondersteunend thuisklimaat.
Gezien de feitelijke toelating van allochtone leerlingen tot het voortgezet onderwijs zullen deze leerlingen niet zelden in de ogen van hun leerkrachten een LBO-perspectief hebben. Het is niet uitgesloten dat leerkrachten uit het basisonderwijs aan leerlingen uit etnische minderheidsgroepen geringere eisen stellen. Dit zal niet opgaan voor alle allochtone leerlingen, maar we moeten er rekening mee houden dat sommige onderwijsdoelstellingen niet geldig worden geacht voor allochtone leerlingen. Dit kan met name een rol spelen bij doelstellingen die doorgaans relatief laat in het onderwijs aan bod komen.
Het is mogelijk dat leerkrachten ook ten aanzien van algemeen aanvaarde doelstellingen bij allochtone leerlingen uitzonderingen maken. Wanneer dit verschijnsel bij onderdelen van het onderwijsaanbod in een bepaald vak optreedt, dan kan het oorzaak van itembias zijn.

2.3 Samenvatting

2.3.1 Samenvatting van de mogelijke determinanten van verschillen in de predictieve validiteit van de Eindtoets Basisonderwijs en het advies basisschool

De volgende mogelijke determinanten van verschillen in de predictieve validiteit zijn van belang.
Als een toets voor twee subgroepen dezelfde norm (regressievergelijking) hanteert om aan te geven welk type voortgezet onderwijs het beste gekozen kan worden, dan wordt het schoolsucces in het voortgezet onderwijs van de subgroep met de laagste toetsscore overschat, van de subgroep met de hoogste score onderschat. Omdat allochtone leerlingen meestal lagere toetsscores hebben dan autochtone leerlingen en omdat toetsen (ook de Eindtoets Basisonderwijs) meestal voor alle subgroepen dezelfde regressievergelijking hanteren om de positie op het extern criterium te schatten, mag verwacht worden, dat toetsen het schoolsucces van allochtone leerlingen overschatten en dat van autochtone leerlingen onderschatten.
Verder kan de transformatie van de beoordeling van het niveau van een leerling (advies basisschool) naar een schaal voor schoolsucces afbreuk doen aan de voorspellende waarde van het advies voor subgroepen. Uit onderzoek volgt de verwachting dat het advies basisschool voor allochtone leerlingen een overschatting geeft van het schoolsucces in het voortgezet onderwijs.

2.3.2 Samenvatting van de potentiële bronnen van itembias

Cummins' (1984a) onderscheid naar 'context-reduced' en 'context-embedded' geeft aanwijzingen voor bronnen van itembias. Het onderscheid heeft betrekking op de mate waarin de taalgebruiker ondersteunende informatie ontvangt van de context waarin het taalgebruik is ingebed. Volgens Cummins (1984a) geeft de dimensie 'context-reduced' en 'context-embedded' geen dichotome presentatie van de werkelijkheid weer, maar gaat het om een glijdende schaal. Toetsitems verschillen in de mate waarin ze een beroep doen op contextmateriaal. In onderzoek naar oorzaken van itembias kan de aandacht uitgaan naar de omvang en naar de aard van het context-materiaal en naar de mate waarin de context allochtone leerlingen bij het beantwoorden van het item ondersteunt.
Bij het opsporen van meer specifieke potentiële bronnen van itembias gaat de aandacht allereerst uit naar linguïstische bronnen van itembias. Het gaat hier in het bijzonder om de elementen van taalvaardigheid waarbij de prestaties van allochtone leerlingen verschillen van die van de autochtone leerlingen. Deze verschillen worden besproken op het niveau van woorden, zinnen en teksten en op het niveau van metalinguïstische vaardigheden.
Potentiële bronnen van itembias op het niveau van woorden zijn:
- de betekenis van woorden;
- woorden met een lage woordfrequentie;
- woorden waarbij de context geen aanwijzingen geeft voor de betekenis van het woord;
- abstracte begrippen;
- ambigue woorden waarbij de context geen aanwijzingen geeft voor de betekenis van het woord.

Potentiële bronnen van itembias op het niveau van zinnen zijn:
- ontkennende zinnen;
- passieve zinnen;
- figuurlijk taalgebruik, specifieke idiomatische uitdrukkingen, metaforen.

Potentiële bronnen van itembias op het niveau van teksten zijn:
- teksten die een groot beroep doen op het geheugen;
- teksten waarvan de inhoud minder plausibel is;
- teksten met moeilijke referenties;
- teksten met ongebruikelijke of onjuiste aanduidingen over de structuur van de tekst.

Potentiële bronnen van itembias op het terrein van metalinguïstische vaardigheden zijn:
- items waarbij grammaticale onjuistheden opgespoord moeten worden;
- items waarbij aandacht voor en controle op het taalgebruik als zodanig een rol spelen.

Verder zijn er ook potentiële culturele bronnen van itembias. Te denken valt hierbij aan verschillen in voorkennis van de onderwerpen die in teksten aan de orde worden gesteld. Allochtone leerlingen kunnen door hun culturele achtergrond minder vertrouwd zijn met teksten die bijvoorbeeld qua inhoud gericht zijn op bekendheid met specifieke elementen van de Nederlandse samenleving.
Een andere potentiële culturele bron van itembias heeft betrekking op de verschillen tussen leerlingen in de mate waarin ze ervaring hebben in het maken van toetsen. Hierdoor kunnen sommige leerlingen minder weten welke taken ze moeten uitvoeren en wat effectieve oplossingsstrategieën zijn.
Verder moeten we er rekening mee houden dat allochtone leerlingen bepaalde elementen van het curriculum minder beheersen, omdat leerkrachten afhankelijk van het toekomstperspectief van de leerlingen meer of minder hoge eisen stellen aan de beheersing van bepaalde vaardigheden. Jungbluth (1985) zegt dat van potentiële VWO-leerlingen meer dan van potentiële LBO-leerlingen gevraagd wordt op het terrein van woordbenoeming, zinsontleding, hanteren van tabellen en grafieken, spelling en geschiedenis. Veel allochtone leerlingen stromen door naar het LBO en MAVO. Het is mogelijk dat leerkrachten aan allochtone leerlingen in dit verband minder hoge eisen stellen. Het is nog een open vraag in hoeverre deze bron van itembias afbreuk doet aan de constructvaliditeit van de toets. In hoofdstuk acht komen we hierop terug.

3 Beschrijving en verantwoording van de onderzoeksinstrumenten

Ten behoeve van het onderzoek naar toets- en itembias zijn vijf instrumenten gebruikt. Twee instrumenten zijn speciaal voor het onderhavige onderzoek gecontrueerd, drie instrumenten zijn binnen de reguliere cyclische activiteiten van het Cito ontwikkeld. De constructie van de Eindtoets Basisonderwijs en van de twee vragenlijsten voor het verzamelen van toelatings- en doorstroomgegevens in het voortgezet onderwijs behoren tot de cyclische activiteiten van het project Eindtoets Basisonderwijs van het Cito. Voor het verzamelen van achtergrondgegevens op leerling- en schoolniveau zijn speciaal voor dit onderzoek twee vragenlijsten samengesteld, die integraal zijn weergegeven in Bijlage 1 en 2.
De onderzoekspopulaties bestaan uit de leerlingen die in 1987 en 1989 deelnamen aan de Eindtoets Basisonderwijs.
De scholen zijn in november 1986 en 1988 via een speciaal bulletin van het Cito (Cito, 1986c; Cito, 1988a) geïnformeerd over het onderzoek naar toets- en itembias. In dat bulletin werden ze tevens uitgenodigd aan het onderzoek deel te nemen. Op 17 en 19 februari 1987 en op 14 en 16 februari 1989 is de Eindtoets Basisonderwijs afgenomen; in diezelfde periode zijn de bovengenoemde vragenlijsten op leerling- en schoolniveau aan alle deelnemende scholen aangeboden.
In het kader van het toelatings- en doorstroomonderzoek is in mei 1987 en 1989 een vragenlijst naar de basisscholen gezonden over de vraag naar welke school voor voortgezet onderwijs de leerlingen zullen vertrekken. In juni 1988 en 1990 zijn de feitelijk toelatings- en doorstroomgegevens in het voortgezet onderwijs verzameld.

In 3.1 komt de opzet van de Eindtoets Basisonderwijs 1987 en 1989 aan de orde, in 3.2 worden de vragenlijsten op leerling- en schoolniveau verantwoord en in 3.3 wordt ingegaan op de vragenlijsten voor het verzamelen van de toelatings- en doorstroomgegevens.

3.1 Opzet van de Eindtoets Basisonderwijs 1987 en 1989

De Eindtoets Basisonderwijs van het Cito, waarvan elk jaar een nieuwe versie verschijnt, heeft twee functies. Enerzijds verschaft de toets informatie over individuele leerlingen in verband met de overgang naar het voortgezet onderwijs, anderzijds levert de toets informatie ten behoeve van de evaluatie van het onderwijsprogramma van de school.

3.1.1 De inhoud en constructie van de Eindtoets Basisonderwijs

De inhoud van de toets wordt gebaseerd op belangrijke communale doelstellingen: algemeen aanvaarde doelstellingen, die voor alle leerlingen van het basisonderwijs geldig worden geacht (Cito, 1986a). De communale doelstellingen die in de toets aan de orde komen, hebben alleen betrekking op cognitieve vaardigheden op het gebied van taal, rekenen en informatieverwerking (zie tabel 3.1). De toetsinhoud wordt verantwoord in het

Doelenboek, de inhoudsverantwoording van de Eindtoets. Voor de Eindtoets Basisonderwijs uit 1987 en 1989 was de versie van het Doelenboek uit 1986 geldig (Cito, 1986a).
Vooral in verband met de massale deelname aan de toets is machinale verwerking van de antwoorden vereist. De toets bestaat derhalve volledig uit objectief scoorbare opgaven met geprecodeerde antwoorden: 180 vierkeuze-opgaven. De 60 taal- en 60 informatieverwerkingopgaven zijn in de opgavenboekjes verdeeld over drie taken van elk 20 opgaven, de 60 rekenopgaven zijn verdeeld over twee taken van elk 30 opgaven. De taal-, reken-, en informatieverwerkingonderdelen van de Eindtoets Basisonderwijs kunnen weer onderverdeeld worden in opgavenrubrieken. De verdeling van de 180 items over de drie toetsonderdelen Taal, Rekenen en Informatieverwerking en de 18 opgavenrubrieken van de Eindtoets 1987 en 1989 is opgenomen in tabel 3.1. Bovendien geeft deze tabel de verdeling van de opgavenrubrieken over de afname-eenheden (de taken) in de opgavenboekjes aan.

Tabel 3.1 Verdeling van de opgaven van de Eindtoets 1987 en 1989

Onderdeel/opgavenrubriek	Taken	1987	1989
Taal		**60**	**60**
– Correct taalgebruik (excl. spelling)	Taal 1, Taal 2	14	13
– Spellen van werkwoorden	Taal 3	11	10
– Spellen van niet-werkwoorden	Taal 3	9	10
– Interpreteerbaar taalgebruik	Taal 1, Taal 2	18	16
– Passend taalgebruik	Taal 1, Taal 2	4	5
– Inhoud	Taal 1, Taal 2	4	6
Rekenen		**60**	**60**
– Getallen	Rek. 1, Rek. 2	9	7
– Hoofdrekenen	Rek. 1	10	14
– Bewerkingen	Rek. 1, Rek. 2	11	11
– Meten	Rek. 1, Rek. 2	9	6
– Procenten	Rek. 1, Rek. 2	5	3
– Verhoudingen	Rek. 1, Rek. 2	3	4
– Vraagstukken: 1987	Rek. 2	13	
1989	Rek. 1, Rek. 2		15
Informatieverwerking		**60**	**60**
– Hanteren van informatiebronnen	Info. 2	6	6
– Kaartlezen	Info. 2	7	7
– Lezen van tabellen en grafieken	Info. 2	7	7
– Lezen van teksten: reproduktie	Info. 1, Info. 3	17	21
– Lezen van teksten: conclusie	Info. 1, Info. 3	23	19
Totaal		**180**	**180**

De constructieprocedure die voor de toets uit 1987 en 1989 geldt, wordt beschreven in Wijnstra (1984a), Engelen & Uiterwijk (1990) en Uiterwijk & Engelen (1992). In het kort komt het op het volgende neer. De op basis van het Doelenboek (Cito, 1986a) geconstrueerde items worden als proeftoets afgenomen bij een steekproef uit scholen die in het betreffende jaar aan de Eindtoets Basisonderwijs deelgenomen hebben. De steekproef wordt systematisch verdeeld in subgroepen. Elke subgroep, die uit ongeveer 400 leerlingen bestaat, maakt ongeveer 90 proeftoetsopgaven. Het aantal subgroepen is afhankelijk van het aantal geconstrueerde opgaven. De proeftoets wordt enkele weken na de Eindtoets gemaakt. Van de items worden naast p- en a-waarden ook punt-biseriële correlaties (r_{pb}) berekend. Een r_{pb} is te beschouwen als een schatting van de item-testcorrelatie, die aangeeft in hoeverre het item meet wat de totale toets meet. Door de proeftoetsgegevens te koppelen aan het Eindtoetsbestand kan voor elk antwoordalternatief van een item de punt-biseriële correlatie berekend worden met de score op het overeenkomstig onderdeel uit de Eindtoets. De aldus berekende proeftoets-gegevens blijken een goede indicatie te geven van de in de Eindtoets te verwachten p- en r_{ir}- waarden (Engelen & Uiterwijk, 1990; Uiterwijk & Engelen, 1992).
De criteria die gelden voor de samenstelling van de concept-toets, zijn de volgende.

- Het in de toetsspecificatie per doelstellingenrubriek vastgelegde aantal opgaven, waarbij naar een evenredige spreiding over de eventuele subrubrieken moet worden gestreefd.
- De p-waarden mogen een spreiding vertonen van 0.45 – 0.90, met dien verstande dat de gemiddelde p-waarde uitkomt tussen 0.70 en 0.75. De item-restcorrelatie van het goede antwoord moet groter zijn dan 0.25.
- Per opgave moet het percentage foute antwoorden zo gelijkmatig mogelijk verdeeld zijn over de foute antwoordmogelijkheden, waarbij de afleider-restcorrelaties niet positief mogen zijn.

In enkele gevallen worden in de concept-toets om inhoudelijke redenen opgaven opgenomen, waarvan de psychometrische gegevens buiten de nagestreefde grenzen vallen.

Items zijn operationalisaties van een bepaald construct en maken gebruik van bepaalde contexten (zie 2.2). Om differentiële kennis van de contexten zo min mogelijk variantie in de scores te laten opeisen, wordt variatie van context-materiaal nagestreefd. Verder is het de bedoeling dat de gebruikte contexten geen specifieke voorkennis (zie 2.2.3) vereisen.
Het contextmateriaal voor de opgaven behorende tot de taken Taal 1, Taal 2, Informatieverwerking 1 en Informatieverwerking 3 bestaat uit teksten die dienen te variëren qua lengte, moeilijkheid, onderwerp en bron (krant, tijdschrift, boek, enzovoort). De leerkrachten die de proeftoets afnemen, wordt gevraagd aan te geven in hoeverre ze de volgende stellingen over de teksten onderschrijven.

- Het onderwerp dat in de tekst aan de orde wordt gesteld en wat daarover gezegd wordt, spreekt de leerlingen aan.
- De gebruikte woorden zijn in het algemeen van een passende moeilijkheid of worden, voorzover ze onvoldoende bekend zijn, in de tekst begrijpelijk

toegelicht.
- De zinnen zijn in het algemeen van een passend niveau qua lengte en gecompliceerdheid.
- De grote lijn in de tekst is voor leerlingen duidelijk te volgen.
- De inhoud van de tekst zal leerlingen emotioneel niet zo raken dat ze daarvan hinder ondervinden bij het maken van de daaropvolgende opgaven.
- Het is nauwelijks voorstelbaar dat de tekst kwetsend is voor een bepaalde bevolkingsgroep.

Tot slot kunnen de leerkrachten aangeven in welke mate ze de tekst geschikt vinden voor opname in de Eindtoets Basisonderwijs.

De teksten die door de meeste leerkrachten op vrijwel alle punten geschikt bevonden worden, komen in aanmerking voor opname in de Eindtoets Basisonderwijs, indien uiteraard voldoende bijbehorende items aan de gestelde psychometrische eisen voldoen.

Bij de taak Rekenen 1, waarin de hoofdrekenitems zijn opgenomen (zie tabel 3.1), mogen de leerlingen geen uitrekenpapier gebruiken. De items zijn in de taken zoveel mogelijk gegroepeerd per opgavenrubriek. In de taken Taal 1 en Taal 2, Informatieverwerking 1 en Informatieverwerking 3, waar de vragen betrekking hebben op teksten, wordt de volgorde van de items in belangrijke mate bepaald door de tekstinhoud.

De door het Cito samengestelde concept-toets wordt met reserve-opgaven ter beoordeling voorgelegd aan een aantal externe vak- en onderwijsdeskundigen. Na de verwerking van het commentaar van de externe screeners wordt de toets definitief vastgesteld. Indien een item uit de concept-toets wordt vervangen door een reserve-opgave, dan worden opnieuw dezelfde drie criteria aangelegd als bij de samenstelling van de concept-toets.

3.1.2 Schaalconstructie voor de rapportage op leerlingniveau

Om de scores van een toets, die moet functioneren voor de keuze van een school voor voortgezet onderwijs, te kunnen interpreteren, moet een relatie gelegd kunnen worden tussen de scores en de verschillende typen voortgezet onderwijs. Bij de Eindtoets Basisonderwijs gebeurt dit door toelatings- en doorstroomgegevens te gebruiken van leerlingen die in een voorgaand jaar aan de toets deelnamen. Aan de hand van de behaalde totaalscore (geëquivaleerde standaardscore) wordt de positie geschat die de leerling in de verschillende typen voortgezet onderwijs zal innemen als de leerling naar dat type zal gaan (zie figuur 1.1). Deze schatting is gebaseerd op onderzoek naar de score-verdeling in de diverse typen voortgezet onderwijs.

Om de toelatings- en doorstroomgegevens van een vorige generatie in een bepaald jaar te kunnen gebruiken, worden deze gegevens gekoppeld aan de zogenaamde geëquivaleerde standaardscores, die van jaar tot jaar vergelijkbaar zijn. Ten behoeve van deze equivalering wordt elk jaar door de leerlingen uit een steekproef van scholen naast de Eindtoets Basisonderwijs van dat jaar ook een ankertoets gemaakt. Een ankertoets die 45 opgaven bevat, is afgezien van het aantal opgaven, vergelijkbaar met de Eindtoets. De resultaten op de

ankertoets die in voorgaande jaren ook door een steekproef is gemaakt, worden gebruikt in een equivaleringsprocedure die door Angoff (1971) als design IV A is beschreven. In dit design wordt toets x gemaakt door groep A en toets y door groep B. Bij de Eindtoets zijn dit de toetsen, respectievelijk toetsdeelnemers van twee opeenvolgende jaren. De groepen A en B maken eveneens toets z (in dit geval de ankertoets). De standaarddeviatie en het gemiddelde van zowel toets x als y worden door middel van ankertoets z geschat voor de gecombineerde steekproef C, die bestaat uit groep A en B samen. Bij de Eindtoets vindt tenslotte een lineaire transformatie plaats naar de standaard-scoreschaal met een gemiddelde van 535, een standaarddeviatie van 10 en met een range van 501 tot en met 550. Deze schaal wordt gedefinieerd aan de ruwe scoreschaal van een bepaald jaar: de zogenaamde moederschaal. Door dezelfde ankertoets een aantal jaren te gebruiken kan men de scores van opeenvolgende jaren blijven projecteren op deze moederschaal (Wijnstra, 1984a; Van der Sman & Uiterwijk, 1985; Engelen & Uiterwijk 1990; Uiterwijk & Engelen, 1992).
In 1987 en 1989 zijn de scores gebracht op de moederschaal van het jaar 1985. Daardoor konden de toelatings- en doorstroomgegevens van de leerlingen die in 1985 aan de toets deelnamen, gebruikt worden om de standaardscores op de leerlingrapporten van de leerlingen uit 1987 te interpreteren. De toelatings- en doorstroomgegevens van de toetsdeelnemers uit 1987 zijn gebruikt voor de interpretatie van scores op de leerlingrapporten in 1989.

3.2 Verantwoording van de vragenlijsten op leerling- en schoolniveau

Om achtergrondgegevens van de leerlingen en scholen te verzamelen die niet via de reguliere activiteiten van het Cito beschikbaar komen, zijn één vragenlijst op school- en één op leerlingniveau ontwikkeld. Voor de ontwikkeling van beide instrumenten is in 1986 in Breda en omgeving een vooronderzoek gehouden. Van dit vooronderzoek is verslag gedaan in De Jong e.a. (1987). In deze paragraaf worden alleen de beslissingen aan de orde gesteld die van belang zijn voor de interpretatie van de uitkomsten van het onderhavige onderzoek.

Voor onderzoek naar itembias voor leerlingen uit etnische minderheidsgroepen zijn per leerling op zijn minst gegevens nodig over de etnische groep waartoe de leerling behoort of gerekend wordt en over het al of niet goed beantwoorden van de in het geding zijnde toetsitems.
Voor onderzoek naar toetsbias zijn meer achtergrondgegevens van de leerling noodzakelijk. Voor het bepalen van de effecten van determinanten van schoolloopbanen van allochtone en autochtone leerlingen zijn variabelen nodig, waarvan bekend is of waarvan verwacht mag worden, dat ze variantie in schoolsucces van de onderscheiden groepen verklaren. Voor de interpretatie van schoolloopbaanmodellen van allochtone en autochtone leerlingen is het bovendien van belang om te beschikken over variabelen waarvan bekend is of waarvan mag worden aangenomen dat de effecten op de schoolloopbanen bij de beide groepen niet gelijk zijn.

De vragenlijsten op school- en leerlingniveau zijn samen met het toetsmateriaal van de Eindtoets Basisonderwijs 1987, respectievelijk 1989 naar de scholen verzonden en geretourneerd naar het Cito. Door de vragenlijsten naar het Cito

terug te laten zenden voordat de toetsscores beschikbaar zijn, kunnen de antwoorden van de leerkrachten niet beïnvloed zijn door de concrete toetsuitslagen, het oordeel van de leerkracht over een leerling is dus onafhankelijk van de door de leerling behaalde concrete Eindtoetsscore tot stand gekomen.

3.2.1 Vragenlijst op leerlingniveau

De vragenlijst over de leerlingen (zie Bijlage 1) bevat in 1987 zeven en in 1989 zes vragen, die door de leerkrachten van groep acht beantwoord moeten worden. Eén vraag gaat over het herkomstland van de ouders, drie, respectievelijk twee vragen hebben betrekking op de schoolloopbaan van de leerling in het Nederlandse basisonderwijs en bij de laatste drie vragen geeft de leerkracht zijn oordeel over enkele leerlingkenmerken. De antwoorden op de vragen konden de leerkrachten op optisch leesbare antwoordbladen aanstrepen.

a *Land van herkomst*
Bij de constructie van de vragenlijst is met betrekking tot het land van herkomst rekening gehouden met twee mogelijke problemen bij onderzoek naar itembias: variatie in de achtergrondkenmerken van etnische groepen en het minimum aantal leerlingen dat per etnische groep nodig is in verband met de statistische analyses.

In 2.2 is gesteld dat een toetsitem een beroep doet op een aantal samenhangende deelvaardigheden. Bij een partijdig item kan niet altijd met zekerheid worden aangegeven welk element van het item op welke (deel)vaardigheid een beroep doet. Dit maakt het moeilijk om bij partijdige items de oorzaak van itembias te detecteren. Dit probleem wordt groter naarmate de linguïstische en culturele verschillen tussen de te vergelijken etnische groepen diffuser zijn. Met in linguïstisch en cultureel opzicht homogene etnische groepen is het eerder mogelijk om bij de bestudering van partijdige items aanknopingspunten te vinden voor oorzaken van itembias. Om variatie in de achtergrondkenmerken binnen de etnische groepen te beperken, is besloten om een leerling tot een bepaalde etnische groep te rekenen, wanneer beide ouders tot de betreffende groep behoren. Bij éénouder gezinnen geldt de etnische achtergrond van de ouder bij wie het kind woont.

Bij de indeling in etnische groepen moeten relatief kleine etnische groepen noodzakelijkerwijs uitgesloten worden, omdat het aantal leerlingen per groep bij statistisch itemanalyses aan een ondergrens gebonden is. Een exacte ondergrens is evenwel moeilijk te geven. In dit verband zijn de analyses van Intraprasert (1986) richtinggevend. Hieruit blijkt dat bij een groep van 400 – 500 leerlingen de samenhang tussen de indices van alle gebruikte statistische itembiasprocedures het grootst is.

In het onderhavige onderzoek is voor de indeling van de etnische groepen aangesloten bij de vijfdeling van Extra & Vallen (1985): mediterrane landen, ex-koloniale gebieden, China, politieke vluchtelingen en overige landen (vgl. ook Extra & Verhoeven, 1990). Op basis van de schatting van het aantal leerlingen dat in groep acht van het basisonderwijs te verwachten is, zijn de vijf hoofdgroepen opgesplitst in 12 subgroepen.

1 Nederland
2 Turkije
3 Marokko
4 Zuid-Europa (Italië, Spanje, Portugal, Griekenland en het toentertijd nog bestaande Joego-Slavië)
5 Oost-Europa (Polen, Hongarije, Roemenië, Bulgarije, Albanië en de toentertijd nog bestaande landen D.D.R., Tsjecho-Slowakije en U.S.S.R.)
6 Noord- en West-Europa (alle overige Europese landen, exclusief Nederland)
7 China (Volksrepubliek China, Taiwan, Hong Kong, Singapore)
8 Molukken
9 Antillen
10 Suriname: Creolen
11 Suriname: Hindoestanen
12 Overige

Omdat bij Molukkers en Chinezen ook tweede-generatiekinderen voorkomen, wordt een leerling ook tot deze etnische groepen gerekend, wanneer de grootouders uit het betreffende 'land' afkomstig zijn. De groep Surinamers is opgesplitst in twee groepen omdat het aantal leerlingen dit toelaat en omdat de culturele en talige verschillen tussen Creolen en Hindoestanen een onderscheid rechtvaardigen. Bovendien worden betrokken leerkrachten in dit geval in staat geacht onderscheid te kunnen maken tussen Creoolse en Hindoestaanse Surinamers.
De variatie in de groep Overige is erg groot. Enerzijds behoren tot deze groep de kinderen van wie het herkomstland niet onder 1 tot en met 11 genoemd wordt, anderzijds behoren tot deze groep de kinderen van wie de ouders niet hetzelfde herkomstland hebben. De groep Overige wordt in feite gebruikt om de andere groepen in linguïstisch en cultureel opzicht zo homogeen mogelijk te houden, maar is zelf te heterogeen om in het onderzoek naar itembias te kunnen functioneren. Deze groep wordt bij de rapportage van de onderzoeksresultaten dan ook buiten beschouwing gelaten.

Met betrekking tot de gekozen indeling in etnische groepen moet worden opgemerkt dat alle onderscheiden etnische groepen in linguïstisch en cultureel opzicht niet even homogeen zijn. Bij de interpretatie van de onderzoeksresultaten moeten we er rekening mee houden, dat ouders uit dezelfde onderscheiden etnische groep, toch in linguïstisch en cultureel opzicht aanzienlijk van elkaar kunnen verschillen. Deze opmerking geldt vooral voor ouders die afkomstig zijn uit Marokko (Driessen, 1991b), Zuid-, Oost-, Noord- en West-Europa en China.

b *Schoolloopbaan in het Nederlandse onderwijs*
De schoolloopbaan van een leerling kan een indicatie geven van de schoolresultaten. Driessen (1990) geeft aan dat de ervaring van verschillende etnische minderheidsgroepen in het Nederlandse onderwijssysteem sterk verschilt. De totale groep allochtone leerlingen uit zijn onderzoek bestaat voor ruim 30% uit zij-instromers. Van de Chinese leerlingen is ruim de helft pas in de bovenbouw (jaargroep 7 en 8) het Nederlandse basisonderwijs gaan volgen. Van de

Marokkaanse kinderen heeft 40% geen volledig Nederlands basisonderwijs gevolgd; bijna 20% is in de bovenbouw gestart. Wanneer Driessen de Nederlandse onderwijservaring relateert aan taal- en rekenprestaties, dan blijkt dat de lengte van de onderwijservaring vooral samenhangt met hogere prestaties in de Nederlandse taal. Bij de Turkse en Marokkaanse leerlingen correleert de verblijfsduur ook met rekenprestaties.
Uit een onderzoek van Roelandt e.a. (1990) blijkt dat de kans op een gunstige onderwijspositie op 12 – 18-jarige leeftijd voor onderinstromers doorgaans twee keer groter is dan voor zij-instromers. Wanneer onderinstromers zich bovendien sterk oriënteren op de Nederlandse samenleving dan is de kans op een gunstige onderwijspositie drie tot vier keer groter dan voor zij-instromers met een zwakke oriëntatie op de Nederlandse samenleving. Onder zij-instromers verstaan Roelandt e.a. leerlingen die in of na het zevende levensjaar zijn toegetreden tot het Nederlandse onderwijs (1990: 110).
In verband met de effecten die verblijfsduur in het Nederlandse basisonderwijs op het vervolg van de schoolloopbaan kan hebben, is in de voor het onderhavige onderzoek gehanteerde vragenlijst op leerlingniveau een vraag opgenomen over de jaargroep waarin de leerling is gestart. Opgemerkt moet worden dat, zoals in het vooronderzoek is gebleken (De Jong e.a., 1987), de startjaargroep niet bekend kan zijn, wanneer de leerling eerst op een andere Nederlandse school heeft gezeten.

Om de doublurestatus van een leerling te bepalen, is ook gevraagd naar het aantal keren dat de leerling een jaargroep heeft gedoubleerd. Bij allochtone leerlingen is het verband tussen doublurestatus en leeftijd met name voor zij-instromers niet zo groot als bij autochtone, waardoor leeftijd voor allochtone leerlingen een schoolloopbaankenmerk kan zijn. In het schooljaar 1984/1985 verbleef 2% van de vijftienjarige allochtone leerlingen nog in het basisonderwijs (CBS, 1986). Omdat deze groep leerlingen in aantal afneemt, is besloten voorlopig alleen voor de dataverzameling in 1987 een vraag over de leeftijd van de leerling op te nemen. Toen na analyse van de data uit 1987 bleek dat de vraag over de leeftijd geen informatiewaarde had naast de vragen over de startjaargroep en doublurestatus, is deze vraag in 1989 vervallen.

c *Oordeel van de leerkracht over kenmerken van de leerling*
Om de scholen bij het invullen van de vragenlijsten zo min mogelijk te belasten is het aantal vragen over leerlingkenmerken beperkt gehouden. Gepoogd is om vragen over leerlingkenmerken op te nemen, waarvan aangenomen mag worden dat deze een relatief grote bijdrage leveren aan de verklaring van verschillen in schoolloopbanen van autochtone en allochtone leerlingen. Bovendien moeten leerkrachten gezien het vooronderzoek (De Jong e.a., 1987) in staat geacht kunnen worden over een specifiek leerlingkenmerk een oordeel te kunnen geven. Er is gekozen voor een vraag over het cognitief functioneren van de leerling, over het perspectief dat de leerling na de basisschool heeft en over de afstand tussen het sociaal-culturele klimaat op school en thuis.

Uit onderzoek van Jungbluth (1985) blijkt dat het beeld dat leerkrachten uit het basisonderwijs van hun leerlingen hebben in belangrijke mate bepaald wordt door twee dimensies. De items uit zijn vragenlijst blijken na factoranalyse in sterke mate te laden op factoren die betrekking hebben op

– 'de schoolprestatiegeschiktheid' en;
– 'de schoolregimegeschiktheid' van de leerling.
Volgens Jungbluth (1985) kunnen leerkrachten twee relatief onafhankelijke oordelen geven over de vraag: 'hoe slim is de leerling' en 'hoe volgzaam is de leerling'. Voor het onderzoek naar het cognitief functioneren van allochtone en autochtone leerlingen is de dimensie schoolprestatiegeschiktheid van Jungbluth bruikbaar. De items die op de dimensie schoolprestatiegeschiktheid hoog laden worden enerzijds gekenmerkt door oordelen over eigenschappen van de leerling en anderzijds door oordelen over de geschiktheid van de leerling voor een bepaald type voortgezet onderwijs.
Rekening houdend met de mogelijkheid dat leerkrachten vinden dat allochtone leerlingen bij een bepaald prestatieniveau eerder geschikt zijn voor een moeilijker type voortgezet onderwijs dan autochtone leerlingen (Wijnstra, 1984b; De Jong, 1987; Uiterwijk 1990b; Driessen, 1991a), is besloten om naast een vraag over de geschiktheid voor de diverse typen voortgezet onderwijs ook een vraag over het cognitief functioneren van de leerling te stellen. Omdat uit Jungbluth (1985) blijkt dat het item 'groot abstractievermogen' het hoogst laadt op de factor schoolprestatiegeschiktheid, is in de vragenlijst een vraag opgenomen over de mate waarin de leerling volgens de leerkracht beschikt over een groot abstractievermogen.

Bij de vraag naar het oordeel van de leerkracht over de geschiktheid voor de typen voortgezet onderwijs werden de volgende antwoordmogelijkheden gegeven: Speciaal onderwijs (SO of VSO), Basisonderwijs (doubleren groep acht), IBO, IBO/LBO, LBO, LBO/MAVO, MAVO, MAVO/HAVO, HAVO, HAVO/VWO en VWO. Deze in onderzoek veelgebruikte antwoord-mogelijkheden (De Jong, 1987; Kerkhoff, 1988; Driessen, 1990; Boland, 1991; Van Langen & Jungbluth, 1992) zijn geen feitelijke doorstroommogelijkheden. De in de onderwijswerkelijkheid steeds meer voorkomende brede scholen-gemeenschappen, als MAVO/HAVO/VWO, ontbreken immers als alternatief. De antwoordmogelijkheden geven in feite een niveau-indicatie van de leerling weer in termen van de typen voortgezet onderwijs. De combinatie van twee schooltypen vormt een tussencategorie die bruikbaar is wanneer er twijfel is of de leerling geschikt is voor één van beide categoriale typen. Omdat er na groep acht van het basisonderwijs leerlingen zijn die nog een jaar in het basisonderwijs blijven en omdat er leerlingen doorstromen naar het (voortgezet) speciaal onderwijs, zijn ook de antwoordmogelijkheden Speciaal onderwijs en Basisonderwijs toegevoegd.

Voor leerlingen uit etnische minderheidsgroepen kan de mate waarin zij de Nederlandse taal gebruiken, samenhangen met het algemene niveau van de leerprestaties. De leerkracht is vaak niet in staat om over de mate van gebruik van de Nederlandse taal door allochtone leerlingen in algemene zin te oordelen, omdat hij de leerling meestal slechts in de schoolsituatie meemaakt. In onderzoek wordt regelmatig het oordeel van leerkrachten gevraagd over één van de factoren die van invloed zijn op de verwerving van het Nederlands als tweede taal: de sociaal-culturele oriëntatie (De Bot e.a., 1985; Lalleman, 1986; Vermeer, 1986; Roelandt e.a., 1990). Het begrip 'sociaal-culturele oriëntatie' is moeilijk in een enkele vraag of in een beperkt aantal vragen eenduidig te operationaliseren. Vermeer (1986) stelt vast dat de 15 items die hij als

operationalisatie voor sociaal-culturele oriëntatie geconstrueerd heeft, niet erg homogeen zijn. Zijn items vallen uiteen in twee redelijk homogene, maar niet aan elkaar gecorreleerde itemclusters: taalgebruik (taalkeuze van de leerling en van zijn ouders) en motivatie (belangstelling van de leerling en van zijn ouders voor het leren op school). Zijn maat voor sociaal-culturele oriëntatie blijkt, als het om de oriëntatie van de leerling zelf gaat, significant te correleren met de taalvaardigheid Nederlands van de allochtone leerling.
Voor onderzoek naar de determinanten van schoolloopbanen van zowel allochtone als autochtone leerlingen moeten we een maat voor de sociaal-culturele oriëntatie van beide groepen hebben. In verband hiermee is de sociaal-culturele oriëntatie in het onderhavige onderzoek opgevat als de afstand tussen het sociaal-culturele klimaat van de school en thuis. Hoewel de project-medewerkers erkennen met een vaag conceptueel kader te werken, werd in 1987 de leerkracht gevraagd een oordeel uit te spreken over de mate waarin de het sociaal-culturele klimaat op school afwijkt van hetgeen de leerling thuis gewend is.
In de vragenlijst van 1989 is op dit punt nadrukkelijker aangesloten bij het begrip 'onderwijsondersteunend thuisklimaat' zoals Jungbluth (1985) en Driessen (1990) dit gebruiken. Uit het onderzoek van Jungbluth (1985: 94) blijkt dat voor het oordeel van leerkrachten over de prestatie-geschiktheid van leerlingen het onderwijsondersteunend thuisklimaat zelfs iets relevanter is dan het beroepsniveau van de vader. Driessen (1990) geeft aan dat de correlatie tussen sociaal milieu en onderwijsondersteunend thuisklimaat .61 bedraagt. Ook voor allochtone leerlingen lijkt onderwijsondersteunend thuisklimaat een hanteerbaar begrip, omdat uit Driessen (1990: 94) blijkt dat bij de grootste etnische groepen er een sterk lineair verband bestaat tussen de scores op onderwijsondersteunend thuisklimaat en sociaal milieu. In 1989 is daarom de leerkracht gevraagd een oordeel te geven over de vraag of het sociaal-culturele klimaat in het gezin de leerling een goede kans biedt op een succesvolle schoolloopbaan in het Nederlandse onderwijssysteem.

d *Een maat voor sociaal-economische status?*
In onderwijssociologisch onderzoek wordt de sociaal-economische status (SES) van de leerling vaak geoperationaliseerd aan de hand van de beroeps- en opleidingsstatus van de vader en de moeder (Meijnen, 1984; Tesser & Mulder, 1990). Voor de uitwerking van de beroepsstatus wordt veelal gebruik gemaakt van de ITS-beroepenklapper van Westerlaak e.a. (1975), die een indeling naar zes niveaus kent: ongeschoolde arbeid, geschoolde arbeid, lagere employées, kleine zelfstandigen, middelbare employées en hogere beroepen. Voor de opleidingsstatus worden de (diploma's van de) Nederlandse onderwijstypen op enigerlei wijze geschaald.
Kerkhoff (1988) en Meijnen & Riemersma (1992) wijzen erop dat de instrumenten die doorgaans gehanteerd worden voor de operationalisatie van SES van autochtonen in minderhedenonderzoek onbruikbaar zijn. Kerkhoff (1988) geeft aan dat de sociale gelaagdheid in de landen van herkomst lang niet altijd vergelijkbaar is met die in West-Europa. Als allochtonen in de Westeuropese landen op de laagste trede van de sociale ladder behoren, wil dat nog niet zeggen dat ze ook in het land van herkomst de laagste positie vertegenwoordigden.
Verder moet gesteld worden dat het opleidingsniveau van de ouders in het land

van herkomst meestal niet goed te beschrijven is met de categorieën die in Nederland gehanteerd worden voor het classificeren van het opleidingsniveau van autochtonen. Bovendien moeten de schalen die voor beroeps- en opleidingsniveau gehanteerd worden zowel aan de boven- als onderkant van de schaal voldoende differentiëren.
Het operationaliseren van SES aan de hand van de beroeps- en opleidingsstatus van de ouders kent ook een aantal praktische problemen. Tesser, Mulder & Van der Werf (1991) vonden dat gegevens over beroep en vooral over de opleiding van de ouders in veel gevallen niet in de schooladministraties staan vermeld. Zij besloten de oudervragenlijsten door de ouders te laten invullen. Voor de ouders van Turkse en Marokkaanse leerlingen was er een brief in de etnische groepstaal bijgevoegd. De respons bij de ouders van allochtone leerlingen was laag: éénderde deel. De respons bij de Surinaamse ouders lag zelfs op éénvierde deel.

Uit Nederlands onderzoek blijkt overigens dat SES bij het ene onderzoek niet hetzelfde is als bij het andere. Alleen al uit de bijdragen aan het thema Onderwijs en samenleving van de Onderwijs Research Dagen 1990 (Klaassen & Jungbluth, 1990) blijkt een grote diversiteit aan SES-operationalisaties. Jungbluth & Van Langen (1990) hanteren voor de SES van leerlingen het beroep en de opleiding van zowel de vader als de moeder. Mulder & Tesser (1990) gebruiken als SES-maat het beroep van de vader en de opleiding van de vader en moeder. Weide & Van der Werf (1990) hanteren alleen de opleiding van beide ouders. De Lange & Rupp (1990) gebruiken in hun onderzoek in de stad Utrecht een sterk afwijkende SES-maat. De stad is ingedeeld in vierkanten van 100 bij 100 meter. De Lange & Rupp hanteren als SES van leerlingen het gemiddelde beroepsniveau van de mannen die in het vierkant wonen, waar de leerling ook woont (1990: 356).

In het onderhavige onderzoek is lang geworsteld met de vraag of en hoe SES geoperationaliseerd zou moeten worden gelet op het feit dat de leerkrachten de gevraagde informatie moeten verstrekken. Bovendien is er rekening mee gehouden dat leerkrachten uit het basisonderwijs niet altijd bereid zijn om via een schriftelijke enquête informatie over de SES van leerlingen aan derden te verstrekken. In de periode van de vragenlijstconstructie (zomer 1986) was het verstrekken van SES-gegevens in de vorm van de leerlinggewichten voor de formatieregeling Wet op het Basisonderwijs niet voor iedere school vanzelfsprekend. Gezien de verwachte weerstanden bij scholen om deze gegevens te verstrekken en gezien de problemen bij het adequaat operationaliseren van SES bij allochtone leerlingen, is besloten geen vraag over de SES van leerlingen op te nemen. Er is wel besloten om in de vragenlijst op schoolniveau een vraag over de samenstelling van de SES van de schoolbevolking op te nemen.

3.2.2 Vragenlijst op schoolniveau

De vragenlijst op schoolniveau (zie Bijlage 2) bevat twee vragen, waarbij de leerkrachten de antwoorden op de vragenlijst konden invullen. De eerste vraag heeft betrekking op de samenstelling van de SES van de totale schoolbevolking. In 1987 en in 1989 is gevraagd om per leerlinggewicht voor de formatieregeling Wet op het Basisonderwijs het aantal leerlingen te vermelden. Bij de tweede

vraag wordt gevraagd naar het aantal leerlingen uit groep acht dat niet aan de Eindtoets Basisonderwijs heeft deelgenomen, omdat ze de Nederlandse taal onvoldoende beheersen om de opgaven te kunnen lezen. De eerste vraag wordt gebruikt om het effect van de samenstelling van de school naar SES op de schoolloopbanen van allochtone en autochtone leerlingen bij de overgang van basisonderwijs naar voortgezet onderwijs te schatten. Met de tweede vraag wordt per school informatie verkregen over het aantal allochtone en autochtone leerlingen in groep acht. Met dit gegeven kan nagegaan worden of de samenstelling van groep acht naar land van herkomst invloed uitoefent op de schoolloopbanen van allochtone en autochtone leerlingen bij de overgang naar het voortgezet onderwijs.

3.3 Toelatings- en doorstroomonderzoeken

Voor het verzamelen van de toelatings- en doorstroomgegevens in het voortgezet onderwijs zijn twee vragenlijsten ontwikkeld. Eén vragenlijst is bedoeld om bij het basisonderwijs per schoolverlater op te vragen naar welke school voor voortgezet onderwijs de leerling zal vertrekken. Met de tweede vragenlijst wordt het voortgezet onderwijs verzocht aan te geven in welk onderwijstype de leerling oorspronkelijk is geplaatst en welke overgangsbeslissing er aan het einde van het eerste leerjaar is genomen. De constructie, de afname en de verwerking ervan behoren tot de cyclische activiteiten van het project Eindtoets Basisonderwijs van het Cito. De toelatings- en doorstroomgegevens zijn verzameld van de leerlingen die in 1987 of in 1989 aan de Eindtoets Basisonderwijs deelnamen.

In mei 1987 en 1989 zijn de basisscholen gevraagd op optisch leesbare antwoordbladen per schoolverlater aan te geven naar welke school voor voortgezet onderwijs de leerling vertrekt. In beide jaren is eind juni een rappèl verzonden.
In oktober 1987 en 1989 zijn de door de basisscholen opgegeven scholen voor voortgezet onderwijs benaderd met het verzoek te controleren of de leerlingenlijst van hun school correct was. De scholen konden op een mutatielijst aangeven welke op de lijst voorkomende leerlingen niet op de school waren ingeschreven en welke leerlingen op de lijst ontbraken.
De leerlingen die volgens opgave van het basisonderwijs nog in het basisonderwijs blijven of doorstromen naar het (Voortgezet) Speciaal Onderwijs, worden in het toelatings- en doorstroomonderzoek verder buiten beschouwing gelaten, omdat zij bij de rapportage van de Eindtoetsscores niet als categorie gehanteerd worden (vgl. Cito, 1988b; Cito, 1990).

In juni van 1988 en 1990 zijn de scholen voor voortgezet onderwijs aangeschreven met het verzoek om de feitelijke toelatings- en doorstroomgegevens te verstrekken. De scholen konden op optisch leesbare antwoordbladen aangeven in welk type eerste leerjaar de leerling oorspronkelijk was geplaatst. De scholen hadden hierbij in 1990 de keuze uit:
- IBO;
- LBO;
- LBO/MAVO;

- LBO/AVO (alle vormen van LBO/AVO die niet vallen onder LBO/MAVO);
- MAVO;
- MAVO/HAVO;
- MAVO/HAVO/VWO;
- HAVO;
- HAVO/VWO;
- VWO.

Op de vragenlijst uit 1988 was het onderwijstype LBO/MAVO in verband met het te verwachten geringe aantal leerlingen nog niet onderscheiden van het type LBO/AVO.
De scholen voor voortgezet onderwijs werd ook gevraagd te vermelden naar welk type tweede leerjaar de leerling zou doorstromen. Daarbij konden dezelfde onderwijstypen gekozen worden als bij het eerste leerjaar. Voor de leerlingen die zouden doubleren, was op het antwoordblad een aparte aanstreeppositie. In september is een rappèl verzonden. Voor de analyse van de gegevens is het toelatings- en doorstroombestand gekoppeld aan het Eindtoetsbestand. In hoofdstuk vier worden de resultaten van de analyses beschreven.

3.4 Samenvatting

Voor het onderhavige onderzoek zijn vijf instrumenten ontwikkeld. De constructie van de Eindtoets Basisonderwijs en van twee vragenlijsten voor het verzamelen van toelatings- en doorstroomgegevens in het voortgezet onderwijs behoren tot de cyclische activiteiten van het project Eindtoets Basisonderwijs van het Cito. Voor het verzamelen van achtergrondgegevens op leerling- en schoolniveau zijn speciaal voor dit onderzoek twee vragenlijsten ontwikkeld. Deze vragenlijsten zijn als Bijlage 1 en 2 integraal opgenomen. De onderzoekspopulaties bestaan uit de leerlingen van groep acht die in 1987, respectievelijk 1989 aan de Eindtoets Basisonderwijs deelnamen. De scholen is gevraagd in de periode van de toetsafname (halverwege februari) de vragenlijsten op leerling- en schoolniveau in te vullen. In mei na de toetsafname is aan de basisscholen gevraagd aan te geven naar welke school voor voortgezet onderwijs elke schoolverlater gaat. Ongeveer een jaar later is aan de betreffende scholen van voortgezet onderwijs gevraagd in welk type eerste leerjaar de leerling is geplaatst en naar welk type tweede leerjaar de leerling zal doorstromen of dat er sprake is van doubleren.

De Eindtoets Basisonderwijs is een schoolvorderingentoets die bestaat uit 180 vierkeuze-opgaven op het gebied van taal, rekenen en informatieverwerking. Deze drie toetsonderdelen worden elk weer onderverdeeld in opgavenrubrieken. De toetsinhoud wordt verantwoord in het Doelenboek, de inhoudsverantwoording van de Eindtoets Basisonderwijs.
Op het leerlingrapport van de Eindtoets Basisonderwijs wordt naast de scores voor Taal, Rekenen en Informatieverwerking een standaardscore vermeld die door een equivaleringsprocedure van jaar tot jaar vergelijkbaar is. De toelatings- en doorstroomgegevens in het voortgezet onderwijs van een vorige generatie leerlingen worden gekoppeld aan deze standaardscore, waardoor voor de interpretatie van de toetsscores de relatie gelegd kan worden tussen de

standaardscore die een leerling heeft behaald en de standaardscoreverdeling van de leerlingen die naar de onderscheiden typen voortgezet onderwijs zijn gegaan.

De vragenlijst op leerlingniveau bevat vragen over
- het land van herkomst;
- de schoolloopbaan van de leerling in het Nederlandse onderwijs;
- het oordeel van de leerkracht over het abstractieniveau van de leerling;
- het oordeel van de leerkracht over de afstand tussen het sociaal-culturele klimaat op school en thuis;
- het oordeel van de leerkracht over de geschiktheid van de leerling voor de verschillende typen van voortgezet onderwijs.

De vragenlijst op schoolniveau bevat vragen over
- de samenstelling van het sociaal milieu van de gehele schoolbevolking;
- het aantal leerlingen uit groep acht dat niet aan de Eindtoets Basisonderwijs deelneemt, omdat ze de Nederlandse taal onvoldoende beheersen om de opgaven te kunnen lezen.

Met de vragenlijsten van het toelatings- en doorstroomonderzoek is vastgesteld in welk type voortgezet onderwijs de leerling in het eerste en in het tweede leerjaar is geplaatst. Op de vragenlijst konden de leerkrachten uit het voortgezet onderwijs kiezen uit alle in de werkelijkheid voorkomende (combinaties van) typen voortgezet onderwijs.

4 Toetsresultaten en toelatings- en doorstroomgegevens van deelnemers aan de Eindtoets Basisonderwijs 1987 en 1989

Van de leerlingen die in 1987, respectievelijk in 1989 aan de Eindtoets Basisonderwijs deelnamen, zijn drie gegevensbestanden opgebouwd. Het eerste bestand bestaat uit de resultaten van de leerlingen die deelnamen aan de Eindtoets Basisonderwijs van het betreffende jaar. Het tweede bestand is een deelverzameling uit het eerste en het derde bestand is een deelverzameling uit het tweede.
In 1987 en 1989 zijn van respectievelijk 80 685 en 92 448 leerlingen Eindtoetsgegevens verwerkt. Deze data vormen het eerste bestand. Het tweede bestand bestaat uit de gegevens van leerlingen waarvan zowel Eindtoets- als vragenlijstgegevens (zie 3.1 en 3.2) beschikbaar zijn. Na koppeling van de Eindtoets- en vragenlijstbestanden bleken er 59 046, respectievelijk 62 716 leerlingen in beide bestanden voor te komen. Hoewel van veel leerlingen de vragenlijstgegevens op schoolniveau bleken te ontbreken, zijn deze leerlingen – om zoveel mogelijk leerlingen in de analyses te kunnen betrekken – toch niet uit het tweede bestand verwijderd. Het tweede bestand wordt gebruikt voor het onderzoek naar itembias, de basisgegevens worden vermeld in 4.2.
Het derde bestand bestaat uit de leerlingen waarvan naast de Eindtoets- en vragenlijstgegevens, ook toelatings- en doorstroomgegevens in het voortgezet onderwijs (zie 3.3) voorhanden zijn. Het derde bestand wordt gebruikt voor het onderzoek naar toetsbias, de basisgegevens worden vermeld in 4.3.
Van de deelnemers uit 1987 zijn er 38 156 leerlingen die in alle drie de bestanden voorkomen; van de deelnemers uit 1989 geldt dat voor 42 420 kinderen. Het tweede bestand bevat 73.2% en 67.8% van het totaal aantal Eindtoetsdeelnemers uit 1987, respectievelijk 1989. Van alle Eindtoetsdeelnemers uit 1987 ontbreken er 21 639 leerlingen (26.8%) in het tweede bestand, van de deelnemers uit 1989 ontbreken er 29 732 (32.2%). Het derde bestand bevat 47.3% en 45.9% van het totaal aantal Eindtoetsdeelnemers uit 1987, respectievelijk 1989. Van alle Eindtoetsdeelnemers uit 1987 ontbreken er 42 529 leerlingen (52.7%) in het derde bestand, van de deelnemers uit 1989 ontbreken er 50 028 (54.1%).

In 4.1 zal het vraagstuk van representativiteit besproken worden. In 4.2 wordt ingegaan op toetsresultaten van de onderscheiden etnische groepen, terwijl in 4.3 de toelatings- en doorstroomgegevens zijn opgenomen. In 4.4 wordt dit hoofdstuk afgesloten met een samenvatting.

De gegevens in hoofdstuk vier zijn gebaseerd op eenvoudige frequentieverdelingen, kruistabellen en berekening van gemiddelden en standaarddeviaties. In hoofdstuk vijf worden de verschillende variabelen met elkaar in verband gebracht.

4.1 Representativiteit

Er is onderzocht of de leerlingen uit het tweede en derde bestand met betrekking tot de toetsresultaten representatief geacht mogen worden voor alle

Eindtoetsdeelnemers uit 1987, respectievelijk 1989. Vooraf moet opgemerkt worden dat de betekenis van representativiteit ten aanzien van de Eindtoets Basisonderwijs beperkt is. De populatie die bestaat uit de leerlingen van scholen die in een bepaald jaar aan de Eindtoets deelnemen, verandert van jaar tot jaar. Zo deden er in 1989 bijvoorbeeld 887 scholen meer mee aan de toets dan in 1987. Dit heeft tot gevolg dat de representativiteit van een steekproef alleen geldt voor de populatie van het betreffende jaar.
Toch moet gesteld worden dat de toetsinhoud en het algemene prestatieniveau van de populaties van jaar tot jaar sterk vergelijkbaar zijn. De verdeling van de opgaven over de opgavenrubrieken van de Eindtoets 1987 en 1989 komt sterk overeen (zie tabel 3.1). Bovendien verschilt de moeilijkheidsgraad van de toetsen voor de populatie van 1987 en 1989 nauwelijks. De gemiddelde p-waarden van de drie toetsonderdelen en totaal liggen in beide jaren tussen .70 en .75. De p-waarden van beide toetsen zijn in aanzienlijke mate te vergelijken, omdat het gemiddelde prestatieniveau van de beide populaties nauwelijks verschilt. De gemiddelden van de geëquivaleerde standaardscores (zie 3.1.2) van alle toetsdeelnemers in 1987 en 1989 verschillen slechts 0.15 standaard-scorepunt (laagste standaarddeviatie is 9.65 standaardscorepunt).
Verder moet opgemerkt worden dat uitkomsten van onderzoek naar significante verschillen tussen grote groepen leerlingen voorzichtig geïnterpreteerd moeten worden, omdat een groot aantal waarnemingen al gauw tot significante verschillen leidt.

In de eerste plaats is onderzocht of de leerlingen uit het tweede bestand, de leerlingen met Eindtoets- en vragenlijstgegevens, met betrekking tot de toetsresultaten representatief geacht mogen worden voor alle Eindtoets-deelnemers uit 1987, respectievelijk 1989. Hiervoor zijn de frequentie-verdelingen van de ruwe scores van de respondenten uit het tweede bestand op de drie toetsonderdelen en de totale toets met de χ^2-toets ('goodness of fit') onderzocht op afwijkingen ten opzichte van de frequentieverdelingen van alle Eindtoetsdeelnemers. Ruwe scores met een frequentie van 4 of lager zijn samengevoegd tot een score-interval met minstens vijf waarnemingen.
De resultaten van deze toetsingen zijn opgenomen in tabel 4.1.

Tabel 4.1 Representativiteit van het tweede bestand voor alle Eindtoetsdeelnemers 1987 resp. 1989 (n.s. = niet significant)

Onderdeel	χ^2	df	p
Eindtoets 1987 (n = 59 046)			
Taal	12.40	50	n.s.
Rekenen	20.46	55	n.s.
Informatieverwerking	16.25	49	n.s.
Totaal	51.31	144	n.s.
Eindtoets 1989 (n = 62 716)			
Taal	16.51	43	n.s.
Rekenen	15.84	50	n.s.
Informatieverwerking	10.46	42	n.s.
Totaal	34.46	109	n.s.

Uit tabel 4.1 blijkt dat het tweede bestand met betrekking tot de frequentieverdeling van de ruwe scores op de drie toetsonderdelen en de totale toets representatief geacht mag worden voor alle Eindtoetsdeelnemers. Gezien het grote aantal leerlingen moeten deze verdelingen in sterke mate gelijkvormig zijn. Omdat de groep autochtone leerlingen onder de respondenten erg groot is (1987: 54 358; 1989: 62 716), is hieruit om praktische redenen een a-selecte steekproef van omstreeks 5 000 leerlingen getrokken. Deze steekproeven bleken met betrekking tot de frequentieverdelingen van de ruwe scores representatief voor alle autochtone leerlingen.

Ook bij het derde bestand, de leerlingen met naast Eindtoets- en vragenlijstgegevens ook toelatings- en doorstroomgegevens, is nagegaan of deze leerlingen representatief geacht mogen worden voor alle Eindtoetsdeelnemers in 1987, respectievelijk 1989. Hiertoe zijn de frequentieverdelingen van de ruwe scores van de respondenten op de drie toetsonderdelen en de totale toets met de χ^2-toets ('goodness of fit') onderzocht op afwijkingen ten opzichte van de frequentieverdelingen van alle Eindtoetsdeelnemers. Ruwe scores met een frequentie van 4 of lager zijn samengevoegd tot een score-interval met minstens vijf waarnemingen. De resultaten van deze toetsingen zijn opgenomen in tabel 4.2.

Tabel 4.2 Representativiteit van het derde bestand voor alle Eindtoetsdeelnemers 1987 resp. 1989

Onderdeel	χ^2	df	p
Eindtoets 1987 (n = 38 156)			
Taal	328.59	50	<.001
Rekenen	294.35	55	<.001
Informatieverwerking	255.43	49	<.001
Totaal	383.20	144	<.001
Eindtoets 1989 (n = 42 420)			
Taal	280.81	43	<.001
Rekenen	280.67	50	<.001
Informatieverwerking	256.90	42	<.001
Totaal	373.08	109	<.001

Uit tabel 4.2 volgt dat het derde bestand als geheel niet representatief geacht mag worden voor alle Eindtoetsdeelnemers. Verwacht mag worden dat het grote aantal waarnemingen van invloed is op de resultaten van deze toetsingen, omdat uit nadere analyse is gebleken dat de curven van de onderzochte verdelingen in hoge mate gelijkvormig zijn.
De autochtone leerlingen in het derde bestand zijn bij vrijwel alle toetsingen representatief voor alle autochtone leerlingen in het tweede bestand. Alleen de ruwe-scoreverdeling van het toetsonderdeel Rekenen uit 1989 van de autochtone leerlingen uit de derde bestand wijkt significant af ($p<.05$) van die van de autochtone leerlingen uit het tweede bestand.
Het tweede bestand dat representatief geacht mag worden voor alle toets-

deelnemers van dat jaar, wordt gebruikt voor het onderzoek naar itembias. Het derde bestand wordt gebruikt voor het onderzoek naar toetsbias. De uitkomsten van de analyses naar toetsbias kunnen, zoals tabel 4.2 laat zien, strikt genomen niet gelden voor alle toetsdeelnemers van het betreffende jaar. In verband met de interpretatie van deze uitkomsten moet er wel op gewezen worden dat het algemene prestatieniveau van de beide derde bestanden in zeer grote mate vergelijkbaar is: de gemiddelde geëquivaleerde standaardscores verschillen slechts 0.01 punt (laagste standaarddeviatie is 9.34 standaardscorepunt). Hierdoor zijn de toelatings- en doorstroomgegevens en de gegevens over toetsbias, in zoverre ze betrekking hebben op het algemene prestatieniveau van de respondenten, in aanzienlijke mate vergelijkbaar. Verschillen tussen de gegevens uit 1987 en 1989 kunnen niet of nauwelijks toegeschreven worden aan een ongelijk gemiddeld prestatieniveau van de beide derde bestanden.
Er kan geen antwoord gegeven worden op de vraag of het prestatieniveau van alle Eindtoetsdeelnemers in een bepaald jaar representatief geacht mag worden voor dat van alle leerlingen in groep acht van het basisonderwijs, omdat er geen geschikte gegevens beschikbaar zijn om deze relatie te bepalen.

4.2 Toetsresultaten van de deelnemers aan de Eindtoets Basisonderwijs 1987 en 1989

Het algemene prestatieniveau van de onderscheiden etnische groepen wordt uitgedrukt in de gemiddelde geëquivaleerde standaardscore. In tabel 4.3 staan per etnische groep het aantal waarnemingen, de gemiddelde standaardscore en de standaarddeviatie op de Eindtoets Basisonderwijs 1987 en 1989. De leerlingen van Nederlandse herkomst vormen een steekproef uit het totaal aantal autochtone leerlingen.

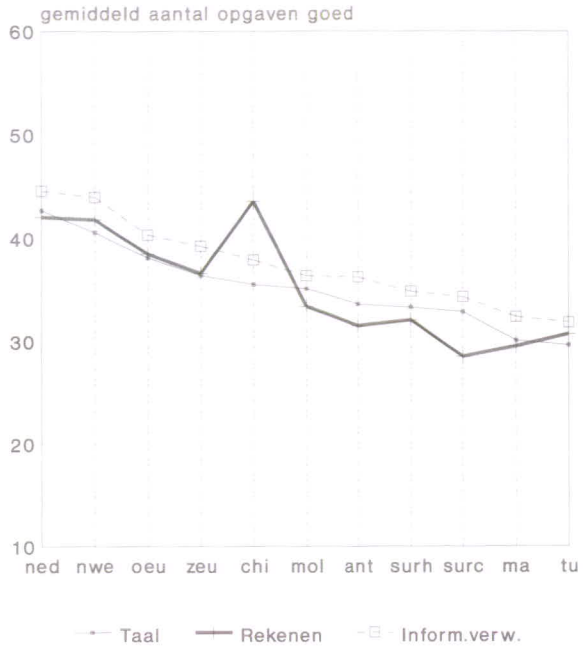
Tabel 4.3 Per etnische groep het aantal leerlingen, de gemiddelde standaardscore en de standaarddeviatie op de Eindtoets 1987 resp. 1989

	Eindtoets 1987			Eindtoets 1989		
Groep	n	M	sd	n	M	sd
Nederland	4969	535.46	9.15	5000	535.81	9.58
Noord- en West-Europa	153	535.06	9.78	150	535.52	10.26
China	150	531.97	10.18	187	533.80	8.81
Oost-Europa	45	531.96	10.58	39	535.32	8.89
Zuid-Europa	209	530.40	10.28	238	529.60	9.95
Molukken	215	527.98	9.55	220	529.69	9.52
Antillen	104	526.84	10.32	161	526.11	10.57
Suriname: Hindoestanen	334	526.46	9.30	294	526.90	10.19
Suriname: Creolen	391	524.98	9.56	377	525.20	10.06
Turkije	797	523.81	9.67	919	523.32	10.21
Marokko	720	523.76	9.32	907	523.73	10.19
	η^2=.19	p<.001		η^2=.20	p<.001	

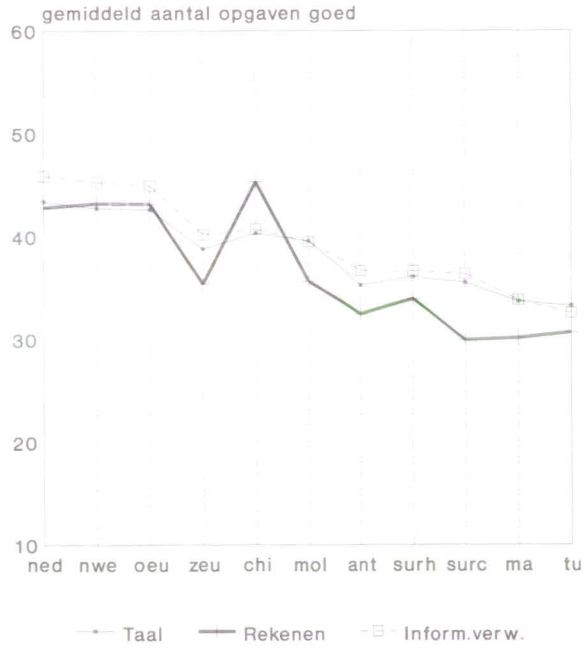
De gemiddelde standaardscores van de onderscheiden etnische groepen op de beide toetsen verschillen in de meeste gevallen nauwelijks. De verschillen tussen 1987 en 1989 bij de leerlingen met ouders uit Oost-Europa kunnen voor een deel ook veroorzaakt worden door het geringe aantal waarnemingen.
De gemiddelde standaardscores van de onderscheiden etnische groepen binnen een bepaald jaar verschillen significant ($p<.001$). Van de totale variantie in de standaardscores wordt 19, respectievelijk 20% (η^2 of eta^2 x 100) verklaard door de factor etnische groep. De gemiddelde score van elke etnische minderheidsgroep is ook paarsgewijs getoetst ten opzichte van de gemiddelde score van de autochtone leerlingen. De t-toetsen voor onafhankelijke steekproeven geven aan dat de scores van de leerlingen uit Noordwest-Europa en Oost-Europa in beide jaren niet significant afwijken van die van de autochtone leerlingen. De standaardscores van de overige etnische minderheidsgroepen wijken wel significant af ($p<.001$; China 1989: $p<.01$). Hierbij moet opgemerkt worden dat het verschillend aantal leerlingen per groep van invloed kan zijn op de resultaten van de toetsingen. Dit geldt met name voor de groep Oost-Europa. Gemiddeld hebben de leerlingen van Marokkaanse en Turkse herkomst de meeste moeite met de 180 Eindtoetsopgaven. Hun gemiddelde standaardscore ligt ongeveer 1.2 standaarddeviatie onder het gemiddelde van de Nederlandse leerlingen. De leerlingen van wie de beide ouders afkomstig zijn uit Suriname, de Antillen of de Molukken en die als groep meer vertrouwd zijn met de Nederlandse taal, scoren in de meeste gevallen toch nog ongeveer één standaarddeviatie onder het gemiddelde van de autochtone leerlingen. Uit ander onderzoek (Driessen, 1990; Mulder, 1993) is eveneens gebleken dat Marokkaanse en Turkse leerlingen meestal de laagste gemiddelde toetsscores behalen en dat de Surinaamse, Antilliaanse en Molukse leerlingen qua gemiddelde toetsscores doorgaans een positie tussen de autochtone en de Marokkaanse/Turkse leerlingen innemen.

De inhoud van de Eindtoets Basisonderwijs is onderverdeeld in Taal, Rekenen en Informatieverwerking. In figuur 4.1 en 4.2 worden de gemiddelde ruwe scores van de onderscheiden etnische groepen op deze drie onderdelen van de Eindtoets 1987 en 1989 gepresenteerd. De ordening van de etnische groepen in de figuren 4.1 en 4.2 is gebaseerd op de gemiddelde taalscore die de onderscheiden groepen in 1987 behaalden. Deze figuren maken per toetsonderdeel de vergelijking tussen de etnische groepen mogelijk. De gemiddelde scores van de etnische groepen van 1987 en 1989 zijn niet volledig vergelijkbaar, omdat de onderdeelscores niet geëquivaleerd worden (Engelen & Uiterwijk, 1990; Uiterwijk & Engelen, 1992).

Figuur 4.1 Resultaten per etnische groep op de Eindtoets Basisonderwijs 1987



Figuur 4.2 Resultaten per etnische groep op de Eindtoets Basisonderwijs 1989



	1987		1989	
Taal	η^2=.19	p<.001	η^2=.16	p<.001
Rekenen	η^2=.13	p<.001	η^2=.15	p<.001
Informatieverwerking	η^2=.18	p<.001	η^2=.21	p<.001

Toelichting:

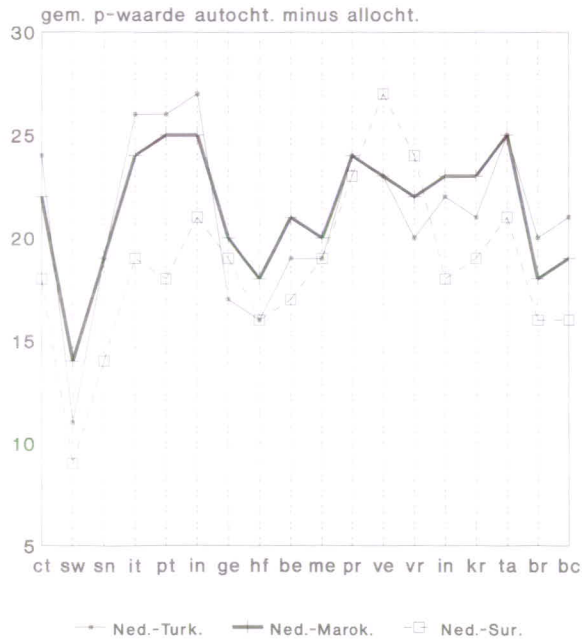
ned = *Nederland*
nwe = *Noord- en West-Europa*
oeu = *Oost-Europa*
zeu = *Zuid-Europa*
chi = *China*
mol = *Molukken*
ant = *Antillen*
surh = *Suriname: Hindoestaan*
surc = *Suriname: Creool*
ma = *Marokko*
tu = *Turkije*

De gemiddelde ruwe scores Taal, Rekenen en Informatieverwerking van de onderscheiden etnische groepen binnen een bepaald jaar verschillen significant (p<.001). De scores van elke etnische minderheidsgroep zijn ook paarsgewijs getoetst ten opzichte van de scores van de autochtone leerlingen. De t-toetsen voor onafhankelijke steekproeven geven aan dat de scores op Taal, Rekenen en Informatieverwerking van de leerlingen uit Noordwest-Europa en Oost-Europa in beide jaren niet significant afwijken van die van de autochtone leerlingen. De rekenscores van de Chinese leerlingen wijken in 1987 niet significant af van die van de autochtone leerlingen, in 1989 is dat wel het geval (p<.01). Alle scores van de overige etnische minderheidsgroepen wijken significant af (p<.001) van die van de autochtone leerlingen. Ook hier moet opgemerkt worden dat de resultaten van de vergelijking van Oosteuropese en Nederlandse leerlingen door steekproeffluctuaties kunnen worden beïnvloed.
Uit figuur 4.1 en 4.2 blijkt dat de gemiddelde rekenresultaten van de Chinese leerlingen opmerkelijk zijn: zij behalen zowel in 1987 als in 1989 vergeleken met alle andere etnische groepen (inclusief autochtone leerlingen) de hoogste gemiddelde score. Zoals gezegd zijn de verschillen met de autochtone leerlingen in 1989 zelfs significant (p<.01). In het onderzoek van Driessen (1990) behalen de Chinese leerlingen ook de hoogste gemiddelde rekenresultaten. De Jong (1987) vindt dat op de numerieke subtest van de door hem gebruikte GALO-test, de Chinese leerlingen de hoogste gemiddelde resultaten behalen. In de Verenigde Staten zien we eveneens dat Chinese leerlingen hogere resultaten behalen op reken-/wiskundetoetsen dan alle andere etnische groepen, inclusief de blanke leerlingen (Dossey e.a., 1988). Applebee, Langer & Mullis (1986) en Baratz-Snowden & Duran (1987) vonden zelfs dat Chinese leerlingen die thuis Engels spreken, ook een betere leesvaardigheid hebben dan blanke leerlingen die thuis Engels spreken.
De taalvaardigheid Nederlands van de Turkse en Marokkaanse leerlingen is in het onderhavige onderzoek vergeleken met die van de andere etnische groepen in beide jaren het laagst. De gemiddelde scores van deze twee groepen op Informatieverwerking versterken dit beeld, omdat het onderdeel Informatieverwerking voor tweederde deel uit begrijpend lezen bestaat. De lage toetsresultaten van de Turkse en Marokkaanse leerlingen inzake de beheersing van het Nederlands worden bevestigd door de resultaten van Landelijke Evaluatie van het Onderwijsvoorrangsbeleid (Tesser, Mulder & Van der Werf, 1991;

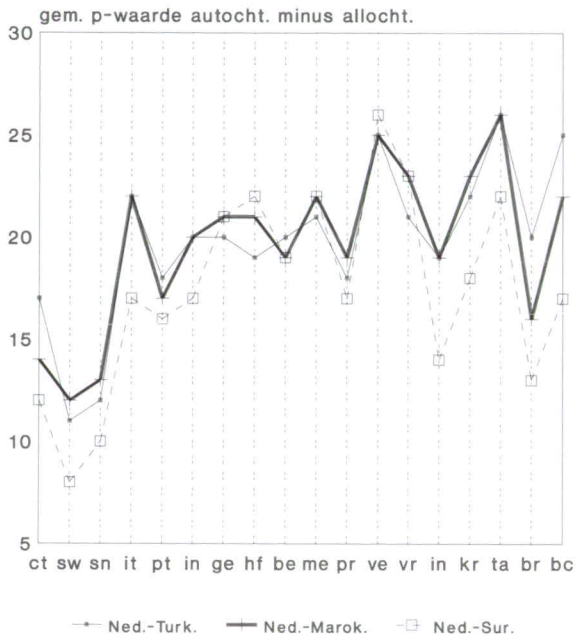
Tesser & Vierke, 1991; Mulder, 1993).

De onderdelen Taal, Rekenen en Informatieverwerking van de Eindtoets Basisonderwijs worden weer onderverdeeld in opgavenrubrieken. Het aantal opgaven per opgavenrubriek is in een aantal gevallen gering (zie toelichting tabellen 4.3 en 4.4). Om de resultaten van de berekeningen niet te sterk door steekproeffluctuaties te laten beïnvloeden, zijn de resultaten per opgavenrubriek alleen van de grootste etnische groepen berekend. Door grotere groepen leerlingen te nemen, blijven steekproeffluctuaties relatief klein, maar eventuele problemen met de inhoudsvaliditeit worden hiermee niet opgelost. Bij de interpretatie van de resultaten van domeinen met een gering aantal items is voorzichtigheid geboden, omdat een meting met meer en/of andere items tot andere resultaten kan leiden.In de figuren 4.3 en 4.4 staan de verschillen tussen de scores van de leerlingen met herkomstland Nederland en met herkomstland Turkije, respectievelijk Marokko en Suriname. De verschilscore is berekend door de gemiddelde score van een bepaalde etnische minderheidsgroep af te trekken van de gemiddelde score van de autochtone leerlingen. De Hindoestaanse en Creoolse leerlingen uit Suriname zijn hier samengenomen in verband met het geringe aantal waarnemingen per afzonderlijke groep.

Figuur 4.3 Verschillen tussen autochtone en allochtone leerlingen per opgavenrubriek van Eindtoets Basisonderwijs 1987



Figuur 4.4 Verschillen tussen autochtone en allochtone leerlingen per opgavenrubriek van Eindtoets Basisonderwijs 1989



Toelichting:

	aantal opgaven	
	1987	*1989*
Taal	***60***	***60***
ct = correct taalgebruik (excl. spelling)	*14*	*13*
sw = spellen van werkwoorden	*11*	*10*
sn = spellen van niet-werkwoorden	*9*	*10*
it = interpreteerbaar taalgebruik	*18*	*16*
pt = passend taalgebruik	*4*	*5*
in = inhoud	*4*	*6*
Rekenen	***60***	***60***
ge = getallen	*9*	*7*
hf = hoofdrekenen	*10*	*14*
be = bewerkingen	*11*	*11*
me = meten	*9*	*6*
pr = procenten	*5*	*3*
ve = verhoudingen	*3*	*4*
vr = vraagstukken	*13*	*15*
Informatieverwerking	***60***	***60***
in = hanteren van informatiebronnen	*6*	*6*
kr = kaartlezen	*7*	*7*
ta = lezen van tabellen en grafieken	*7*	*7*
br = begrijpend lezen: reproduktie	*17*	*21*
bc = begrijpend lezen: conclusies	*23*	*19*

De figuren 4.3 en 4.4 laten zien dat de leerlingen uit de betreffende etnische minderheidsgroepen op alle opgavenrubrieken lager scoren. De gemiddelde scores van elke etnische minderheidsgroep zijn paarsgewijs getoetst ten opzichte van de gemiddelde scores van de autochtone leerlingen. De t-toetsen voor onafhankelijke steekproeven geven aan dat de scores in alle gevallen significant ($p<.001$) afwijken van die van de autochtone leerlingen. De verschillen tussen de gemiddelde scores van autochtone en allochtone leerlingen zijn het kleinst bij de beide spellingrubrieken, met name bij de spelling van de werkwoordvormen. In hoofdstuk zeven wordt hierop nader ingegaan. De Surinaamse leerlingen behalen lage resultaten bij de rekenrubriek Verhoudingen. Gezien het geringe aantal opgaven van deze rubriek (in 1987: 3 opgaven; in 1989: 4 opgaven) is voorzichtigheid bij de interpretatie geboden. Wel kan opgemerkt worden dat de verschillen bij Verhoudingen in beide jaren relatief groot zijn. In beide jaren behalen zowel de Turkse als Marokkaanse leerlingen relatief lage gemiddelde scores bij Lezen van Tabellen en Grafieken. De Turkse leerlingen hebben meer moeite met begrijpend lezen dan hun Marokkaanse en Surinaamse medeleerlingen.

4.3 Toelatings- en doorstroomgegevens van de deelnemers aan de Eindtoets Basisonderwijs 1987 en 1989

Voor het analyseren van de toelatings- en doorstroomgegevens is het derde bestand gebruikt. In 4.1 hebben we geconstateerd dat het derde bestand uit 1987 en uit 1989 niet representatief geacht mag worden voor alle Eindtoetsdeelnemers van dat jaar. Daarbij werd opgemerkt dat de curven van de onderzochte ruwe scoreverdeling van steekproef en populatie echter in hoge mate gelijkvormig zijn. Verder is gezegd dat de gemiddelde geëquivaleerde standaardscores van het derde bestand in beide jaren vrijwel gelijk zijn, waardoor verschillen tussen de toelatings- en doorstroomgegevens 1987 en 1989 niet of nauwelijks zijn toe te schrijven aan verschillen in prestatieniveau.
In tabel 4.4 wordt de verdeling van de leerlingen over het eerste leerjaar van de diverse typen voortgezet onderwijs gegeven (in rijpercentages). Met LBO/AVO worden alle scholengemeenschappen met een LBO- en één of meer AVO-element(en) bedoeld. M/H/V omvat scholengemeenschappen voor MAVO/HAVO of MAVO/HAVO/VWO. HAVO/VWO staat voor HAVO, VWO en scholengemeenschappen voor HAVO/VWO. In tabel 4.4 staan de gegevens van de (steekproef uit) autochtone, alle allochtone leerlingen en van de drie grootste etnische minderheidsgroepen. Omdat er meer etnische minderheidsgroepen zijn dan de drie in tabel 4.4 genoemde, correspondeert het aantal leerlingen bij 'Allochtonen' niet met de som van het aantal Surinaamse, Turkse en Marokkaanse leerlingen.
Er is nagegaan of de verschillen tussen de autochtone en allochtone leerlingen in de verdeling over het eerste leerjaar van de schooltypen van het voortgezet onderwijs significant zijn. Deze berekeningen zijn uitgevoerd op de ruwe data van het derde bestand uit 1987, respectievelijk uit 1989. Aan de schooltypen van het voortgezet onderwijs (zie 3.3) zijn ten behoeve van deze toetsingen numerieke waarden toegekend van 1 (=IBO) tot en met 9 (=VWO), respectievelijk van 1 (=IBO) tot en met 10 (=VWO) (vgl. Bosker, 1990; Van der Velden, 1991). De verschillen tussen de autochtone en allochtone leerlingen in

de verdeling over het eerste leerjaar van de onderscheiden typen voortgezet onderwijs zijn op significantie onderzocht met de t-toets voor onafhankelijke steekproeven. De η^2 (x 100) in tabel 4.4 geeft het percentage variantie in de schooltypen aan dat verklaard wordt door de factor etnische groep. De etnische groepen zijn hier de leerlingen waarvan de ouders afkomstig zijn uit Nederland, Suriname, Turkije en Marokko. De toetsingen voor de tabellen 4.5 – 4.11 zijn op dezelfde wijze uitgevoerd als voor tabel 4.4.

Tabel 4.4 Verdeling van de leerlingen uit 1987 resp. 1989 over het eerste jaar van het voortgezet onderwijs (in rijpercentages)

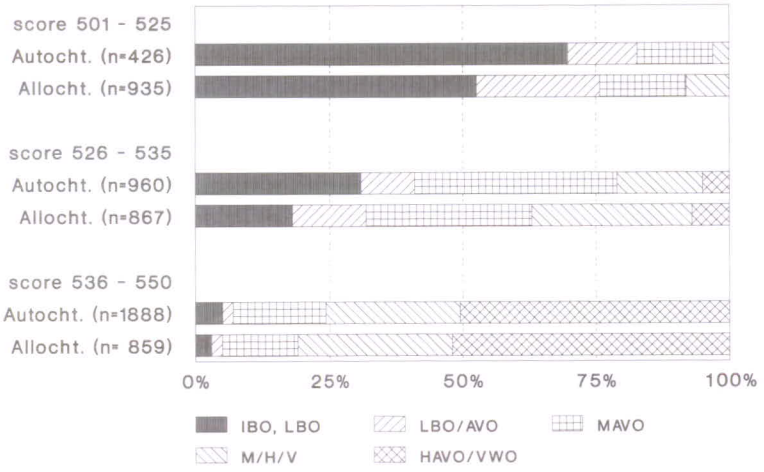
Groep	IBO, LBO	LBO/ AVO	MAVO	M/H/V	HAVO/ VWO	n	t-toets resp. η^2
1987							
Autochtonen	21	6	23	19	31	3274	t=9.71
Allochtonen	25	13	20	22	19	2661	p<.001
Suriname	22	24	16	26	12	386	η^2=.06
Turkije	35	16	23	18	9	431	p<.001
Marokko	40	14	21	20	5	375	
1989							
Autochtonen	19	7	19	23	33	3405	t=10.21
Allochtonen	23	9	22	26	20	3092	p<.001
Suriname	21	13	23	26	17	353	η^2=.07
Turkije	33	12	26	21	9	534	p<.001
Marokko	35	12	26	20	6	540	

Uit tabel 4.4 blijkt dat allochtone leerlingen in vergelijking met autochtone in beide jaren oververtegenwoordigd zijn in IBO, LBO en in scholengemeenschappen voor LBO/AVO, MAVO/HAVO en MAVO/HAVO/VWO. Autochtone leerlingen stromen in beide jaren naar verhouding meer door naar de moeilijker geachte schooltypen: HAVO, VWO en HAVO/VWO. Voor het MAVO verschilt het beeld in beide jaren: in 1987 gaan naar verhouding meer autochtone leerlingen naar het MAVO, in 1989 meer allochtone.
Als we de drie onderscheiden etnische minderheidsgroepen bezien, dan komt naar voren dat de Surinaamse leerlingen naar verhouding meer toegelaten worden tot HAVO, HAVO/VWO, VWO en brede scholengemeenschappen (LBO/AVO, MAVO/HAVO en MAVO/HAVO/VWO). Turkse en Marokkaanse leerlingen treffen we vooral aan in IBO, LBO en MAVO.

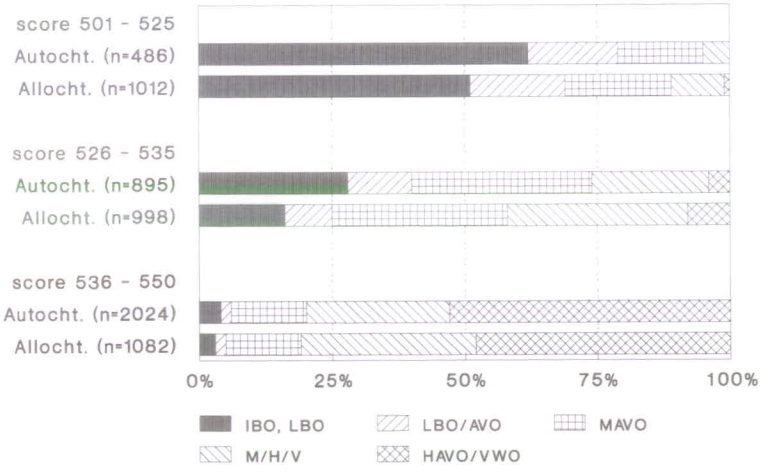
Tabel 4.4 geeft aan wat de onderwijspositie van de onderscheiden etnische groepen is in het eerste leerjaar van het voortgezet onderwijs. Dit beeld wordt over het algemeen bevestigd door andere onderzoeken (vgl. Driessen, 1990; Mulder & Pijl, 1992).
Het verschil tussen de gemiddelde toetsscores van de etnische groepen (zie tabel 4.3) is aanleiding om de toelatings- en doorstroomgegevens van leerlingen

van een vergelijkbaar prestatieniveau te analyseren. Door de leerlingen te groeperen naar prestatieniveau wordt het beeld van de onderwijspositie in de eerste klas van het voortgezet onderwijs verfijnd. De leerlingen toegelaten tot het eerste leerjaar voortgezet onderwijs zijn aan de hand van de Cito-standaardscore ingedeeld in drie score-intervallen: 501 – 525, 526 – 535, 536 – 550 (figuur 4.5 en 4.6). Er is voor deze score-intervallen gekozen, omdat hierdoor de allochtone leerlingen in drie ongeveer even grote groepen worden verdeeld.

Figuur 4.5 Toelatingsgegevens van autochtone en allochtone leerlingen per score-interval Eindtoets Basisonderwijs 1987



Figuur 4.6 Toelatingsgegevens van autochtone en allochtone leerlingen per score-interval Eindtoets Basisonderwijs 1989



Wanneer we naar de leerlingen van een vergelijkbaar prestatieniveau kijken, dan leveren de gegevens over de toelating tot het eerste leerjaar voortgezet onderwijs zowel in 1987 als in 1989 een ander beeld op dan dat van tabel 4.4. Met name uit de twee laagste score-intervallen blijkt dat de onderwijspositie van allochtone leerlingen gunstiger is dan die van autochtone leerlingen: ze gaan meer naar de moeilijker geachte schooltypen dan hun vergelijkbaar presterende autochtone klasgenoten. De gegevens uit de figuren 4.5 en 4.6 zijn relevant genoeg om gedetailleerder te analyseren. De tabellen 4.5 – 4.7 zijn gebaseerd op dezelfde gegevens als de figuren 4.5 en 4.6. Toegevoegd zijn de uitgesplitste gegevens van de drie grootste etnische minderheidsgroepen (Surinaamse, Turkse en Marokkaanse leerlingen). Bij het interpreteren van deze laatste gegevens moet gelet worden op het aantal waarnemingen, omdat één procent soms betrekking heeft op een beperkt aantal leerlingen.

Tabel 4.5 Verdeling van de leerlingen uit standaardscore-interval 501 – 525 uit 1987 resp. 1989 over het eerste jaar van het voortgezet onderwijs (in rijpercentages)

Groep	IBO, LBO	LBO/ AVO	MAVO	M/H/V	HAVO/ VWO	n	t-toets resp. η^2
1987							
Autochtonen	69	13	14	3	0	426	t=4.73
Allochtonen	52	23	16	8	0	935	p<.001
Suriname	38	37	10	15	0	184	η^2=.03
Turkije	54	22	16	7	0	232	p<.001
Marokko	61	18	13	8	0	195	
1989							
Autochtonen	62	17	16	5	0	486	t=4.84
Allochtonen	51	18	20	10	1	1012	p<.001
Suriname	39	23	27	10	1	140	η^2=.02
Turkije	52	14	24	8	1	278	p<.001
Marokko	58	19	15	8	1	264	

Tabel 4.6 Verdeling van de leerlingen uit standaardscore-interval 526 – 535 uit 1987 resp. 1989 over het eerste jaar van het voortgezet onderwijs (in rijpercentages)

Groep	IBO, LBO	LBO/ AVO	MAVO	M/H/V	HAVO/ VWO	n	t-toets resp. η^2
1987							
Autochtonen	31	10	38	16	5	960	t=7.81
Allochtonen	18	14	31	30	7	867	p<.001
Suriname	11	19	23	37	11	123	η^2=.04
Turkije	15	11	35	31	7	141	p<.001
Marokko	24	15	34	23	4	125	
1989							
Autochtonen	28	12	34	22	4	895	t=7.68
Allochtonen	16	9	33	34	8	998	p<.001
Suriname	14	9	26	40	11	125	η^2=.03
Turkije	17	12	33	31	8	159	p<.001
Marokko	19	9	39	29	5	178	

Tabel 4.7 Verdeling van de leerlingen uit standaardscore-interval 536 – 550 uit 1987 resp. 1989 over het eerste jaar van het voortgezet onderwijs (in rijpercentages)

Groep	IBO, LBO	LBO/ AVO	MAVO	M/H/V	HAVO/ VWO	n	t-toets resp. η^2
1987							
Autochtonen	5	2	17	25	50	1888	t=2.34
Allochtonen	3	2	14	29	52	859	p<.05
Suriname	3	3	16	37	42	79	η^2=.00
Turkije	5	2	17	28	48	58	n.s.
Marokko	4	0	18	56	22	55	
1989							
Autochtonen	4	2	14	27	53	2024	t=.24
Allochtonen	3	2	14	33	48	1082	n.s.
Suriname	3	2	13	33	49	88	η^2=.01
Turkije	2	3	24	40	31	97	p<.001
Marokko	3	2	32	39	24	98	

Uit de tabellen 4.5 – 4.7 volgt dat autochtone leerlingen in alle score-intervallen oververtegenwoordigd zijn in IBO en LBO, terwijl in MAVO/HAVO en MAVO/HAVO/VWO en meestal in HAVO, VWO en HAVO/VWO naar verhouding meer leerlingen uit etnische minderheidsgroepen zitten. Het

MAVO laat op dit punt een wisselend beeld zien: in het lage scoregebied zijn de allochtone leerlingen oververtegenwoordigd, in het midden en hoge scoregebied zijn de autochtone leerlingen in de meerderheid of houden beide groepen elkaar in evenwicht. In het HAVO/VWO zijn in de meeste score-intervallen de allochtone leerlingen oververtegenwoordigd. In het hoogste score-interval zijn de verschillen tussen de onderscheiden etnische groepen niet altijd significant. Als we kijken naar de ontwikkeling van 1987 naar 1989, dan vinden we dat in het LBO/AVO in de twee laagste score-intervallen de toestroom van allochtone leerlingen afneemt. In het MAVO neemt het aandeel van de autochtone leerlingen uit de twee hoogste scoregebieden af. Ook is zichtbaar dat er minder autochtone leerlingen naar het IBO en LBO gaan.

De gegevens in de tabellen 4.5 tot en met 4.7 over de leerlingen van Surinaamse, Turkse of Marokkaanse herkomst hebben betrekking op relatief geringe aantallen, waardoor de conclusies minder stellig kunnen zijn dan ten aanzien van de totale groep allochtone leerlingen. Om die reden blijven de gegevens uit tabel 4.7 en de vergelijking 1987 en 1989 bij de nu volgende bespreking geheel buiten beschouwing.
Leerlingen van Surinaamse herkomst treffen we naar verhouding het minst aan op het IBO en LBO, de Marokkaanse leerlingen het meest. De Surinaamse leerlingen blijken bij de toelating tot het voortgezet onderwijs een sterke voorkeur te hebben voor brede scholengemeenschappen. Zij zijn in de meeste score-intervallen oververtegenwoordigd in LBO/AVO, MAVO/HAVO en MAVO/HAVO/VWO. Surinaamse leerlingen gaan doorgaans minder dan vergelijkbaar presterende Turkse en Marokkaanse leerlingen naar het MAVO.

Uit de tabellen 4.5 – 4.7 volgt dat over het algemeen leerlingen uit etnische minderheidsgroepen meer dan hun autochtone klasgenoten toegelaten worden tot schooltypen waar het gemiddelde prestatieniveau hoger ligt. Voor leerlingen van Marokkaanse herkomst geldt dit in mindere mate. Over het algemeen blijken allochtone leerlingen een voorkeur te hebben voor brede scholengemeenschappen, dit geldt in het bijzonder voor Surinaamse leerlingen. In dit soort scholen is het mogelijk om de definitieve schoolkeuze één of meer jaren uit te stellen.

Om een goed beeld van de doorstroming aan het einde van het eerste leerjaar van allochtone en autochtone leerlingen van een vergelijkbaar prestatieniveau te krijgen, is de doorstroming vanuit een aantal schooltypen geanalyseerd. Leerlingen zijn in de nu volgende tabellen dus vergelijkbaar in die zin dat het prestatieniveau van de leerlingen aan het einde van het basisonderwijs aanleiding gaf ze naar hetzelfde schooltype te laten gaan. Vanwege het geringe aantal waarnemingen is de groep allochtone leerlingen hier niet onderverdeeld. In de tabellen 4.8 – 4.11 worden alleen de percentages gegeven van de leerlingen die blijven zitten of doorstromen naar een onderwijstype waar het oorspronkelijk gekozen type niet in voorkomt. Deze groep leerlingen is relevant, omdat het om leerlingen gaat van wie reeds na één jaar bekend is dat het oorspronkelijk gekozen schooltype om de één of andere reden niet de juiste is of dat daaraan op zijn minst getwijfeld kan worden (doublure). Onder elke tabel wordt aangegeven wat vanuit een bepaald schooltype beschouwd wordt als 'afstroom' en wat als 'opstroom'.

Tabel 4.8 Percentage leerlingen uit 1987, resp. 1989 dat vanuit het eerste jaar LBO op- of afstroomt

Groep	Lager	Doublure	Afstroom	Opstroom	n	t-toets
1987						
Autochtonen	1.6	0.8	2.4	0.7	607	t=.47
Allochtonen	6.5	2.7	9.2	1.0	522	n.s.
1989						
Autochtonen	2.3	3.4	5.7	0.2	533	t=.69
Allochtonen	3.8	8.0	11.8	0.9	549	n.s.

Toelichting: lager = IBO; afstroom = lager + doublure; opstroom = MAVO, HAVO, VWO en alle combinaties zonder IBO, LBO

Tabel 4.9 Percentage leerlingen uit 1987, resp. 1989 dat vanuit het eerste jaar MAVO op- of afstroomt

Groep	Lager	Doublure	Afstroom	Opstroom	n	t-toets
1987						
Autochtonen	1.2	1.7	2.9	2.0	750	t=.28
Allochtonen	3.7	4.8	8.5	2.8	539	n.s.
1989						
Autochtonen	3.5	7.3	10.8	1.4	658	t=1.55
Allochtonen	3.9	13.3	17.2	3.5	693	n.s.

Toelichting: lager = IBO, LBO; afstroom = lager + doublure; opstroom = HAVO, VWO en HAVO/VWO

Tabel 4.10 Percentage leerlingen uit 1987, resp. 1989 dat vanuit het eerste jaar MAVO/HAVO en MAVO/HAVO/VWO afstroomt

Groep	Lager	Doublure	Afstroom	n	t-toets
1987					
Autochtonen	2.7	2.1	4.8	632	t=2.93
Allochtonen	1.5	4.9	6.4	586	p<.01
1989					
Autochtonen	1.8	8.1	9.9	770	t=.42
Allochtonen	2.4	12.4	14.8	799	n.s.

Toelichting: lager = IBO, LBO; afstroom = lager + doublure

Tabel 4.11 Percentage leerlingen uit 1987, resp. 1989 dat vanuit het eerste jaar HAVO, HAVO/VWO en VWO afstroomt

Groep	Lager	Doublure	Afstroom	n	t-toets
1987					
Autochtonen	5.4	1.6	7.0	1002	t=.55
Allochtonen	7.4	1.6	9.0	512	n.s.
1989					
Autochtonen	5.4	2.3	7.7	1111	t=1.45
Allochtonen	8.4	5.4	13.8	609	n.s.

Toelichting: lager = IBO, LBO, MAVO; afstroom = lager + doublure

Uit de tabellen 4.8 tot en met 4.11 volgt dat de afstroom van allochtone leerlingen in alle onderzochte schooltypen en in beide jaren hoger is dan die van hun autochtone klasgenoten. De afstroom van allochtone leerlingen is afhankelijk van schooltype en jaar 1.3 tot 3.8 keer groter dan die van autochtone leerlingen. Ook de opstroom in het LBO en MAVO is bij allochtone leerlingen groter. De gegevens in deze tabellen wekken de indruk dat de trefzekerheid waarmee de schoolkeuze van allochtone leerlingen gemaakt wordt, lager is dan die van autochtone leerlingen: de af- en opstroom van allochtone leerlingen is groter. Opgemerkt moet worden dat de verschillen tussen autochtone en allochtone leerlingen vrijwel nergens significant zijn. Hoewel de verschillen dus kennelijk niet groot zijn, zijn ze toch ook niet zonder betekenis: de gegevens uit 1987 en 1989 wijzen systematisch in dezelfde richting. Bij vergelijking van de gegevens uit 1987 en 1989 blijkt dat de afstroom in alle onderzochte onderwijstypen in absolute zin toeneemt. In het LBO en MAVO neemt het aandeel van de allochtone leerlingen in de afstroom af, terwijl dit aandeel van allochtone leerlingen in het MAVO/HAVO, MAVO/HAVO/VWO, HAVO, HAVO/VWO en VWO in geringe mate toeneemt. In het LBO en MAVO wordt in 1987 de afstroom vooral veroorzaakt door de doorstroming naar een onderwijstype met een lager gemiddeld prestatieniveau; in 1989 door doublure. In MAVO/HAVO en MAVO/HAVO/VWO wordt de afstroom zowel in 1987 als in 1989 vooral beïnvloed door het percentage zittenblijvers. Dit geldt in 1989 ook voor de doorstroming vanuit HAVO, HAVO/VWO en VWO.

In de tabellen 4.8 – 4.11 wordt uitsluitend naar de af- en opstroom gekeken. Van belang zijn ook de verschillen tussen autochtone en allochtone leerlingen bij de doorstroming binnen scholengemeenschappen. Hoe is bijvoorbeeld de verhouding tussen beide groepen leerlingen bij de doorstroming vanuit de eerste klas MAVO/HAVO, wanneer de tweede klas bestaat uit klassen voor MAVO en HAVO? Om dit te onderzoeken zijn de verschillen tussen autochtone en allochtone leerlingen met betrekking tot de doorstroming naar de verschillende typen tweede leerjaar vanuit het eerste leerjaar getoetst met de t-toets voor onafhankelijke steekproeven.
Hieruit blijkt dat bij alle soorten scholengemeenschappen de verschillen tussen autochtone en allochtone leerlingen bij de doorstroming vanuit het eerste

leerjaar niet significant zijn. Hoewel de verschillen tussen autochtone en allochtone leerlingen dus niet groot zijn, zijn er wel systematische verschillen. Bij de doorstroming vanuit LBO/(M)AVO zien we zowel in 1987 als in 1989 meer allochtone dan autochtone leerlingen naar een gecombineerde klas voor LBO/(M)AVO gaan. In zoverre er aan het einde van het eerste leerjaar gekozen wordt tussen LBO en MAVO, gaan allochtone leerlingen naar verhouding meer naar LBO en autochtone leerlingen meer naar MAVO. Bij de doorstroming vanuit MAVO/HAVO zien we zowel in 1987 als in 1989 meer allochtone dan autochtone leerlingen naar een gecombineerde klas voor MAVO/HAVO gaan. In zoverre er gekozen wordt tussen MAVO en HAVO, gaan allochtone leerlingen naar verhouding meer naar HAVO en autochtone leerlingen meer naar MAVO. Bij de doorstroming vanuit MAVO/HAVO/VWO zien we dat allochtone leerlingen eerder doorstromen naar MAVO, autochtone leerlingen meer naar HAVO of HAVO/VWO. Bij de doorstroming vanuit HAVO/VWO zijn de verschillen kleiner en minder systematisch.
Hoewel de verschillen tussen autochtone en allochtone leerlingen bij de doorstroming vanuit het eerste leerjaar bij alle soorten scholengemeenschappen niet significant zijn, gaan allochtone leerlingen zowel in 1987 als in 1989 doorgaans naar de schooltypen met een lager gemiddeld prestatieniveau. Het tegengestelde beeld laat het MAVO/HAVO zien.

4.4 Samenvatting

In hoofdstuk vier komen de eerste en tweede onderzoeksvraag aan de orde.

1 *Hoe ontwikkelen de Eindtoetsscores van allochtone en autochtone leerlingen zich van 1987 tot 1989?*

2 *Hoe verloopt de toelating en doorstroming van allochtone en autochtone leerlingen in het voortgezet onderwijs?*

De gemiddelde geëquivaleerde standaardscores van de onderscheiden etnische groepen op de Eindtoets Basisonderwijs 1987 en 1989 verschillen nauwelijks. De leerlingen van Marokkaanse en Turkse herkomst behalen in beide jaren de laagste gemiddelde standaardscore, gevolgd door de leerlingen van Surinaamse, Antilliaanse en Molukse herkomst. De Chinese leerlingen behalen vergeleken met alle andere etnische groepen (inclusief autochtone leerlingen) de hoogste gemiddelde rekenscore. Dit wordt bevestigd door ander onderzoek in Nederland en de Verenigde Staten. De verschillen tussen de gemiddelde scores van autochtone en allochtone leerlingen zijn bij de spellingopgaven, met name bij de spelling van werkwoordvormen, relatief klein.

Wanneer we bij de toelating tot het voortgezet onderwijs de instroom van leerlingen van een vergelijkbaar prestatieniveau bezien dan blijkt dat er meer autochtone dan allochtone leerlingen naar schooltypen met een lager gemiddeld prestatieniveau (IBO, LBO en LBO/AVO) gaan. Allochtone leerlingen hebben over het algemeen de overhand in het AVO in vergelijking met vergelijkbaar presterende autochtone leerlingen. Deze bevindingen komen overeen met die van De Jong (1987), Driessen (1991a) en Van Langen & Jungbluth (1992):

allochtone leerlingen worden tot een hoger schooltype toegelaten dan op grond van toets- of testscores verwacht mag worden.

De relatieve voorsprong die allochtone leerlingen bij de start in het voortgezet onderwijs hebben, wordt volgens de gegevens uit het onderhavige onderzoek voor een deel weer teniet gedaan aan het einde van het eerste leerjaar. Uit de doorstroomgegevens van leerlingen van een vergelijkbaar prestatieniveau blijkt dat allochtone leerlingen meer dan hun autochtone klasgenoten afstromen (blijven zitten of doorstromen naar een onderwijstype met een lager gemiddeld prestatieniveau). Ook Mulder & Tesser (1991) rapporteren dat de afstroom van allochtone leerlingen aan het einde van het eerste leerjaar groter is dan van autochtone leerlingen. Ook blijkt dat allochtone leerlingen toegelaten tot LBO en MAVO, iets meer dan autochtone leerlingen doorstromen naar onderwijstypen met een hoger gemiddeld prestatieniveau. Opgemerkt moet worden dat de verschillen in af- en opstroom zowel in 1987 als in 1989 voorkomen, maar vrijwel nergens significant zijn.
De verschillen tussen autochtone en allochtone leerlingen bij de doorstroming binnen scholengemeenschappen zijn eveneens geanalyseerd. Er is geconstateerd dat bij alle soorten scholengemeenschappen de verschillen tussen autochtone en allochtone leerlingen bij de doorstroming vanuit het eerste leerjaar niet significant zijn. Hoewel de afzonderlijke verschillen niet significant zijn, zijn er wel systematische verschillen: allochtone leerlingen gaan zowel in 1987 als in 1989 doorgaans naar de schooltypen met een lager gemiddeld prestatieniveau, het tegengestelde beeld laat het MAVO/HAVO zien.

Het lijkt erop dat bij de overgang naar het voortgezet onderwijs de schoolkeuze van allochtone leerlingen met een geringere trefzekerheid wordt gemaakt dan die van autochtone leerlingen. In hoofdstuk vijf wordt nagegaan hoe hoog de voorspellende waarde van het advies basisschool en van de Eindtoets Basisonderwijs is voor leerlingen uit etnische minderheidsgroepen.

5 Toetsbias in de Eindtoets Basisonderwijs 1987 en 1989

Toetsen worden onder andere gebruikt om voorspellingen te doen over buiten de toetssituatie liggend gedrag. Op grond van de behaalde toetsscore spreken we dan een verwachting uit over iemands niveau op een bepaald criterium op basis van eerder verworven kennis over de relatie tussen de toetsscores en het criteriumgedrag. Zo wordt in Nederland aan het einde van het basisonderwijs met toetsen de vaardigheid van leerlingen in bijvoorbeeld taal en rekenen bepaald. Met de behaalde toetsscores wordt tevens aangegeven welk schooltype van het voortgezet onderwijs bij de schoolkeuze het meest voor de hand ligt. De taal-, reken- en totaalscores van leerlingen kunnen gebruikt worden om een indicatie te geven van het naar verwachting te behalen niveau van voortgezet onderwijs, omdat uit eerder onderzoek de relatie tussen de toetsscores en de behaalde niveaus van voortgezet onderwijs (het extern criterium) empirisch is vastgesteld.
De relatie toetsscores – criteriumgedrag kan ook voor onderscheiden subgroepen bepaald worden. Er is sprake van toetsbias, wanneer we systematisch schattingsfouten maken bij het voorspellen van de positie op een extern criterium als functie van het groepslidmaatschap (Reynolds, 1982; Malpass & Poortinga, 1986). In feite is een toets onpartijdig wanneer de regressielijnen van een toets op het extern criterium van twee subgroepen samenvallen. Een toets kan op toetsbias onderzocht worden, wanneer van het extern criterium een valide en onpartijdige operationalisatie beschikbaar is (Cronbach, 1972; Jensen, 1980; Reynolds, 1982; Kok, 1988).

In 1.2.1 is gesteld dat het strikt genomen onmogelijk is om te beoordelen of er bij de Cito-Eindtoets Basisonderwijs wel of niet sprake is van toetsbias, omdat er in feite geen maat voor succes in het voortgezet onderwijs beschikbaar is, waarvan de onpartijdigheid is aangetoond.
Aan de andere kant moeten we echter vaststellen dat toetsen en het advies basisschool in de onderwijspraktijk een belangrijke functie vervullen bij de schoolkeuze en bij de toelating tot het voortgezet onderwijs. Daardoor functioneert het onderwijsniveau dat een leerling na een bepaalde periode in het voortgezet onderwijs heeft bereikt, feitelijk als maat voor schoolsucces. Dit betekent ook dat in de praktijk het bereikte onderwijsniveau wordt gehanteerd om de voorspellende waarde van toetsscores en van het advies basisschool te beoordelen.
In deze dissertatie wordt onderzoek naar toetsbias opgevat als het nagaan van de predictieve validiteit van de Eindtoets Basisonderwijs voor de onderscheiden etnische groepen in vergelijking met die van het advies van de basisschool. Om inzicht in de achtergrond te krijgen van de relatie toetsscore, respectievelijk advies basisschool en extern criterium is ook onderzocht wat de effecten zijn van enkele relevante determinanten van schoolloopbanen op de Eindtoetsscore, advies basisschool en schoolsucces van allochtone en autochtone leerlingen. Opgemerkt wordt dat in het onderhavige onderzoek het advies basisschool onafhankelijk van de toetsscores van de leerlingen tot stand is gekomen, omdat de leerkrachten de vragenlijsten, waarop ze het schoolkeuze-advies moesten

invullen, voor de rapportage van de Eindtoetsscores hebben geretourneerd. Alle analyses in dit hoofdstuk zijn uitgevoerd met het bestand dat bestaat uit leerlingen waarvan naast de Eindtoets- en vragenlijstgegevens ook toelatings- en doorstroomgegevens in het voortgezet onderwijs voorhanden zijn (het derde bestand; zie 3.3 en 4.1).

In 5.1 worden de schalen besproken voor de metingen van de onafhankelijke variabelen. In 5.2 komt de constructie van een schaal voor schoolsucces – de afhankelijke variabele – aan bod. In 5.3 staat voor de onderscheiden etnische groepen de relatie tussen het advies basisschool, respectievelijk toetsscore en de schaal voor schoolsucces centraal. In 5.4 worden de effecten van een aantal belangrijke variabelen op de schoolloopbanen van allochtone en autochtone leerlingen geschat. Het hoofdstuk wordt afgesloten met een samenvatting (5.5).

5.1 Meetniveau van de onafhankelijke variabelen

De predictieve validiteit van de Eindtoets Basisonderwijs en het advies basisschool wordt in dit onderzoek twee keer nagegaan: met de gegevens uit 1987 en met die uit 1989. Naast het advies basisschool en de toetsscore gebruiken we voor het schatten van de effecten van determinanten van schoolloopbanen de gegevens die met behulp van de vragenlijst op leerling-niveau verzameld zijn (zie 3.2.1).
Het meetniveau van deze variabelen kan verschillen. Van de variabelen Cito-score, de jaargroep waarin de leerling in het Nederlandse onderwijs is gestart ('startleerjaar'), het aantal keren dat de leerling in het basisonderwijs heeft gedoubleerd ('doublure') en de leeftijd van de leerling ('leeftijd') wordt aangenomen dat ze minstens op intervalniveau worden gemeten. Van een aantal andere variabelen is het niet zeker of er sprake is van een meting op intervalniveau. Dit geldt voor het oordeel van de leerkracht over de geschiktheid van de leerling voor de verschillende typen van voortgezet onderwijs (het advies), het oordeel over het 'abstractievermogen' van de leerling en het oordeel over het 'thuisklimaat' (zie 3.2.1). Voor de beantwoording van de laatste twee vragen konden de leerkrachten gebruik maken van een vijfpuntsschaal lopend van 'eens' tot 'oneens'. Voor het advies basisschool kon gekozen worden uit de antwoordmogelijkheden: Speciaal Onderwijs, Basisonderwijs, IBO, IBO/LBO, LBO, LBO/MAVO, MAVO, MAVO/HAVO, HAVO, HAVO/VWO en VWO (zie 3.2.1).

Ten behoeve van de analyses is aan elke antwoordmogelijkheid van de drie laatstgenoemde vragen een ranggetal toegekend. Hoewel de ranggetallen zo gekozen zijn dat de afstand tussen twee opeenvolgende antwoordmogelijkheden telkens één is, zijn de geconstrueerde schalen hierdoor nog geen schalen op intervalniveau. Gegeven het ordinale niveau van de variabelen is elke monotone transformatie mogelijk zonder dat de ordening van de klassen verandert. De relatie tussen de variabelen kan echter wel veranderen. Om te kunnen bepalen of de relatie tussen de variabelen bij benadering lineair is, is in navolging van Blok & Saris (1980) en Visser & Voeten (1987) nagegaan of de η-coëfficiënten (eta-coëfficiënten) van de verschillende variabelen veel hoger zijn dan de produkt-moment correlatie-coëfficiënten. In tabel 5.1 staan de beide

coëfficiënten van de onderscheiden variabelen van 1987 en 1989 onder elkaar. Bij de vergelijking van deze coëfficiënten hebben de mintekens geen betekenis, omdat de η-coëfficiënten niet negatief kunnen zijn. In deze tabel is ter informatie ook de variabele 'klas 1' opgenomen. Op de schaal 'klas 1' staan de typen voortgezet onderwijs tot welke de leerlingen daadwerkelijk zijn toegelaten. Het zijn dus niet de onderwijstypen die gehanteerd zijn bij het advies basisschool, maar de typen die in werkelijkheid voorkomen: IBO, LBO, LBO/MAVO (alleen in 1989), LBO/AVO, MAVO, MAVO/HAVO, MAVO/HAVO/VWO, HAVO, HAVO/VWO en VWO (zie 3.3). Conform de procedure in 4.3 zijn aan deze onderwijstypen in de hierboven aangegeven volgorde ranggetallen toegekend, beginnend bij één en met een onderlinge afstand van één. De onderwijstypen zijn in deze volgorde geplaatst, omdat dit overeenstemt met de rangorde van de gemiddelde Cito-standaardscore van de leerlingen in de betreffende onderwijstypen (Cito, 1990: 23).
Tabel 5.1 is gebaseerd op de gegevens van alle allochtone leerlingen en op een steekproef uit de autochtone leerlingen (zie tabel 4.3).

Uit tabel 5.1 blijkt dat de verschillen tussen de produkt-moment correlatie-coëfficiënten en de η-coëfficiënten gering zijn. De verschillen tussen de beide coëfficiënten van de variabelen 'startleerjaar', 'doublure' en 'leeftijd' wijken over het algemeen niet sterk af van die van de variabelen 'abstractievermogen', 'verschil school-thuis', 'advies' en 'klas 1'. Dit onderscheid is van belang omdat van de variabelen 'startleerjaar', 'doublure' en 'leeftijd' aangenomen mag worden dat ze op rationiveau gemeten zijn, hetgeen uiteraard niet van de variabelen 'abstractievermogen', 'thuisklimaat', 'advies' en 'klas 1' gezegd kan worden. De verschillen tussen de beide soorten coëfficiënten stemmen overeen met die van Blok & Saris (1980) en Visser & Voeten (1987) en aangenomen kan worden dat de relatie tussen de variabelen bij benadering lineair is. De coëfficiënten in tabel 5.1 geven derhalve geen aanleiding te veronderstellen dat er betekenisvolle schattingsfouten gemaakt worden wanneer we voor deze data het intervalniveau aanvaarden.

Tabel 5.1 Produkt-moment correlaties (boven) en η-coëfficiënten (onder) tussen de variabelen

	Cito-score	start-leerj.	dou-blure	leef-tijd	abstr. verm.	thuis-klimaat	advies
Eindtoetsdeelnemers 1987							
start-leerjaar	-.205 .233						
doublure	-.296 .332	.117 .156					
leeftijd	-.241 .267	.344 .329	.434 .462				
abstractie-vermogen	.656 .679	-.067 .077	-.279 .302	-.171 .187			
thuis-klimaat	-.435 .436	.251 .264	.205 .205	.277 .293	-.283 .320		
advies	.747 .774	-.106 .129	-.312 .352	-.216 .257	.762 .775	-.365 .390	
klas 1	.726 .740	-.094 .139	-.290 .317	-.212 .239	.670 .674	-.356 .371	.825 .827
Eindtoetsdeelnemers 1989							
start-leerjaar	-.152 .173						
doublure	-.318 .320	.117 .119					
leeftijd (niet 1989)							
abstractie-vermogen	.697 .697	-.036 .045	-.304 .312				
thuis-klimaat	.489 .490	-.159 .164	-.264 .270		.494 .498		
advies	.771 .780	-.081 .100	-.335 .360		.810 .816	.523 .531	
klas 1	.729 .736	-.085 .104	-.330 .340		.701 .704	.466 .473	.834 .837

5.2 De constructie van een schaal voor schoolsucces

Voor onderzoek naar de predictieve validiteit van instrumenten die gebruikt worden om relevante informatie te verschaffen over individuele leerlingen in verband met de overgang naar het voortgezet onderwijs, moet een criterium gekozen worden dat kan dienen als afhankelijke variabele en dat kan gelden als maat voor schoolsucces.
In schoolloopbaanonderzoek bestaat reeds enige traditie ten aanzien van het construeren van een variabele die daartoe kan dienen. Uitgangspunt hierbij is dat de schoolloopbanen van leerlingen in het voortgezet onderwijs verschillen en dat de kenmerken van schoolloopbanen hiërarchisch te ordenen zijn. Het begrip schoolloopbaan wordt gestructureerd door vier kenmerken (Van der Velden, 1991).
De schoolloopbanen van leerlingen worden in de eerste plaats gekenmerkt door het *onderwijsniveau* dat leerlingen bereikt hebben. Het onderwijsniveau heeft betrekking op de indeling in de categoriale schooltypen LBO, MAVO, HAVO en VWO. Soms kan deze indeling in onderwijsniveaus verfijnd worden, wanneer binnen in een schooltype niveaus onderscheiden worden (LBO-A tot LBO-D; MAVO-C en MAVO-D). De schoolloopbanen kunnen in de tweede plaats gekenmerkt worden door de verdeling van leerlingen over de *leerjaren* binnen een schooltype. In het verlengde van het tweede kenmerk ligt de vraag of de leerling aan het einde van een leerjaar overgaat of doubleert, respectievelijk wel of niet het diploma behaalt. Het wel of niet succes hebben aan het einde van een leerjaar is derhalve het derde kenmerk: *prestatie*. Het vierde kenmerk heeft betrekking op de *richting* die een leerling binnen een schooltype kiest. De richting binnen een school voor voortgezet onderwijs heeft betrekking op de keuze van het vakkenpakket. Omdat bij de keuze voor de ene richting binnen een schooltype de leerling tot meer vervolgstudies toegelaten kan worden dan bij de andere, kunnen richtingen gekwantificeerd worden. Interesse voor een vakkenpakket of affiniteit met een bepaald beroep als zodanig vallen buiten deze kwantificering. Het gaat uitsluitend om het tellen van het aantal vervolg-opleidingen dat een bepaald vak of een combinatie van vakken als toelatingseis stelt. Zo zijn er bijvoorbeeld in het schooljaar 1983/1984 27 HBO-opleidingen die VWO-abituriënten toelaten wanneer zij wiskunde I en natuurkunde in het vakkenpakket hebben; terwijl 9 HBO-opleidingen van de VWO-abituriënt natuurkunde en scheikunde eisen (Bosker, 1990; Van der Velden, 1991).

De schoolloopbaan van een leerling in het voortgezet onderwijs kan opgevat worden als de opeenvolging van bepaalde posities in het onderwijssysteem, waarbij de positie bepaald wordt door de 4 genoemde kenmerken. In onderzoek heeft de schoolloopbaan soms ook betrekking op het niveau dat een leerling op een bepaald moment in het onderwijs bereikt heeft (Kreft & De Leeuw, 1986), waarbij het niveau bepaald wordt door de schoolloopbaankenmerken onderwijsniveau, leerjaar en richting. Bij de schoolloopbaan als moment is het kenmerk prestatie verdisconteerd in onderwijsniveau en leerjaar. Bij beide benaderingen zullen de schoolloopbanen geschaald moeten worden.
De onderzoeker zal onder meer beslissingen moeten nemen over de afstand tussen de in zijn onderzoek voorkomende onderwijsposities.
Het schalen van mogelijke schoolloopbanen kan op twee momenten geschieden: a posteriori en a priori (Kreft & De Leeuw, 1986; Van der Velden, 1991).

De onderzoeker kan achteraf, wanneer de data over de schoolloopbanen verzameld zijn, de empirische loopbaanpatronen met een bepaalde techniek schalen. Tesser (1986) en Kreft & De Leeuw (1986) kennen a posteriori door middel van homogeniteitsanalyse aan loopbaanpatronen bepaalde scores toe. Sterk vergelijkbare loopbaanpatronen krijgen weinig verschillende scores, terwijl uiteenlopende patronen sterk verschillende scores krijgen. De schoolloopbaan 'HAVO-1 → MAVO-2 → MAVO-3 → MAVO-4' krijgt bijvoorbeeld een score die dicht in de buurt ligt van de loopbaan 'MAVO-1 → MAVO-2 → MAVO-3 → MAVO-4', omdat de laatste 3 onderwijsposities identiek zijn.
De onderzoeker kan op twee manieren a priori beslissingen nemen over de onderlinge afstand van onderwijsposities: op basis van oordelen van experts en op basis van de eigenschappen van het onderwijsstelsel.
Cremers (1980) laat experts oordelen geven over de rangorde van verschillende onderwijsposities. Zo moeten zij de relatie aangeven tussen HAVO-3 en MAVO-4. Vervolgens worden de verschillende oordelen omgezet in schaalwaarden (Cremers, 1980).
Koopman, Van den Eeden & De Jong (1986), Bosker (1990) en Van der Velden (1991) kwantificeren a priori door de eigenschappen van het onderwijsstelsel te gebruiken. Het Nederlandse onderwijssysteem wordt voorgesteld als een hiërarchisch systeem waarbij de leerlingen aan de onderkant binnenstromen en waarbij de top wordt gevormd door de universiteit. De verschillen tussen de onderwijsposities in het voortgezet onderwijs worden aangegeven door het aantal leerjaren dat nodig is om door te stromen naar de top van dit deel van het onderwijssysteem: leerjaar zes van het VWO. Zo is een leerling in VWO-4 twee jaar van de top verwijderd. Bij de overgang naar een hoger leerjaar binnen hetzelfde onderwijstype komt de leerling één trede dichter bij de top, evenals bij de overgang naar een hoger onderwijstype bij gelijkblijvend leerjaar. Bij doublure binnen hetzelfde type blijft de leerling een jaar langer op dezelfde trede staan. De onderwijstypen binnen hetzelfde leerjaar verschillen onderling één trede, omdat de leerlingen bij de overgang naar een hoger onderwijstype één jaar verliezen. De verschillende onderwijsposities worden door Bosker (1990) en Van der Velden (1991) aan de hand van de kenmerken onderwijsniveau, leerjaar en prestatie als een leerjarenladder voorgesteld. Een verkorte versie van deze leerjarenladder wordt weergegeven in figuur 5.1.

Figuur 5.1 Leerjarenladder voortgezet onderwijs

Ladder	Onderwijsposities					
Top						
10	VWO-6					
9	VWO-5					
8	VWO-4	HAVO-5				
7	VWO-3	HAVO-4				
6	VWO-2	HAVO-3	MAVO-4			
5	VWO-1	HAVO-2	MAVO-3	LBO,C-4		
4		HAVO-1	MAVO-2	LBO,C-3	LBO,AB-4	IBO-4
3			MAVO-1	LBO-2	LBO,AB-3	IBO-3
2				LBO-1		IBO-2
1						IBO-1
Bodem						

In het onderhavige onderzoek moeten de posities die de leerlingen na de overgangsbeslissing aan het einde van het eerste leerjaar voortgezet onderwijs bereikt hebben, gehanteerd worden als maat voor schoolsucces. Het gaat hierbij niet om de schoolloopbaan als de opeenvolging van een reeks posities in het onderwijssysteem. De schoolloopbaan heeft in dit onderzoek betrekking op het prestatieniveau dat een leerling na het eerste leerjaar bereikt heeft (zie 3.3). Het prestatieniveau wordt bepaald door de kenmerken onderwijsniveau en leerjaar (zie Van der Velden, 1991).
Voor het (a posteriori) schalen van de bereikte prestatieniveaus is in dit onderzoek geen gebruik gemaakt van homogeniteitsanalyse (vgl. Kreft & De Leeuw, 1986), omdat het om het prestatieniveau op een bepaald moment gaat en niet om een opeenvolging van onderwijsposities. Er is ook niet gekozen voor een procedure waarbij experts a priori het niveau van verschillende onderwijsposities beoordelen (vgl. Cremers, 1980), omdat hierbij niet duidelijk wordt waarom experts de posities ordenen zoals ze doen. Vanwege de eenvoud en begrijpelijkheid is het aantrekkelijker om een leerjarenladder te hanteren waarbij gebruik gemaakt wordt van de kenmerken van het onderwijsbestel (Koopman, Van den Eeden & De Jong, 1986; Bosker 1990; Van der Velden, 1991). Toch kent ook deze procedure een tweetal bezwaren.
Het eerste bezwaar heeft betrekking op het feit dat het verschil tussen twee onderwijstypen altijd door één schaalpunt wordt aangegeven, terwijl de verschillen tussen de prestatieniveaus van de verschillende onderwijstypen in feite niet gelijk zijn (zie tabel 5.2). Het tweede bezwaar heeft betrekking op het feit dat in dit onderzoek niet volstaan kan worden met een leerjarenladder bestaande uit de categoriale schooltypen, omdat leerlingen ook verblijven in het eerste of tweede leerjaar van een combinatie van schooltypen. Zo is het moeilijk om een waarde toe te kennen aan de tweede klas van een HAVO/VWO-combinatie versus de tweede klas van een MAVO/HAVO/VWO-combinatie, omdat de top van het voortgezet onderwijs, VWO-6, vanuit beide klassen in dezelfde tijd te bereiken is.
Om aan de genoemde bezwaren tegemoet te komen, is besloten om aan elke voorkomende onderwijspositie van het tweede leerjaar de waarde toe te kennen

die overeenkomt met het gemiddelde prestatieniveau van alle leerlingen in dat onderwijstype, zoals gemeten door de Eindtoets Basisonderwijs 1987, respectievelijk 1989. Voor de leerlingen die zijn blijven zitten in het eerste leerjaar van een bepaald schooltype, kan niet de gemiddelde score van de zittenblijvers of de eerstejaars-leerlingen uit dat type genomen worden, omdat de gemiddelde score van hen nauwelijks afwijkt van het gemiddelde van de leerlingen die overgaan naar het tweede leerjaar. Toch moet de gemiddelde score van zittenblijvers substantieel afwijken van die van de tweedejaars, omdat de zittenblijvers één jaar vertraging hebben opgelopen. Bij de leerjarenladder (Bosker, 1990; Van der Velden, 1991) is de afstand tussen de zittenblijvers en de leerlingen die overgaan naar het tweede jaar gelijk aan de afstand tussen twee schooltypen: één punt (zie figuur 5.1). Omdat in werkelijkheid het verschil tussen de gemiddelde scores van de schooltypen niet steeds gelijk is (zie tabel 5.2), is het gemiddelde van de verschillen tussen de schooltypen genomen om de afstand tussen twee schooltypen te bepalen. In navolging van de leerjarenladder van Bosker (1990) en Van der Velden (1991) is besloten de leerlingen die doubleren de gemiddelde score van de eerstejaars toe te kennen minus de gemiddelde afstand tussen twee schooltypen. In tabel 5.2 staan de schaalwaarden van de onderwijsposities van de leerlingen uit 1987 en 1989.

Tabel 5.2 Schaalwaarden van de bereikte onderwijsniveaus (afgerond)

	1987		1989	
Type	Doublure	2e leerjaar	Doublure	2e leerjaar
IBO	514.2	517.6	513.2	516.4
LBO/MAVO (alleen 1989)			523.5	526.7
LBO	523.7	527.2	523.7	526.9
LBO/AVO	523.3	526.7	525.7	528.9
MAVO	531.1	534.5	531.2	534.5
MAVO/HAVO	533.1	536.5	533.2	536.4
MAVO/HAVO/VWO	534.1	537.6	535.3	538.5
HAVO	537.2	540.6	537.8	541.1
HAVO/VWO	539.4	542.8	540.0	543.2
VWO	541.7	545.1	542.2	545.4

5.3 De predictieve validiteit van het advies basisschool en de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen

Wanneer leerlingen aan het einde van de basisschool een school voor voortgezet onderwijs moeten kiezen, worden in het schoolkeuzeproces het advies basisschool en in de regel een toets- of testuitslag gebruikt. Het is niet vanzelfsprekend dat de waarde van het advies en de toets- of testscore voor de schoolkeuze van allochtone en autochtone leerlingen vergelijkbaar is.
Om de predictieve validiteit van de Eindtoets Basisonderwijs, respectievelijk het advies basisschool te bepalen, is in de eerste plaats de correlatie berekend tussen de Eindtoets Basisonderwijs, respectievelijk het advies basisschool en de

schaal voor schoolsucces (zie 5.2). Het gaat hier dus telkens om de correlatie tussen één onafhankelijke en één afhankelijke variabele, onafhankelijk van multivariate verbanden.
In tabel 5.3 zijn voor elke onderscheiden etnische minderheidsgroep (zie 3.2.1) de produkt-moment correlatie-coëfficiënten opgenomen van Eindtoets-standaardscore, respectievelijk advies basisschool met de schaal voor schoolsucces.

Tabel 5.3 Produkt-moment correlatie-coëfficiënten tussen de Cito-score, het advies basisschool en de schaal voor schoolsucces in 1987 resp. 1989 *

Groep	Pmc-coëfficiënten 1987			Pmc-coëfficiënten 1989		
	n	Cito-score	advies-bao	n	Cito-score	advies-bao
Autochtonen	3274	.76	.83	3405	.78	.85
Alle allochtonen	2661	.76	.79	3092	.74	.81
Noordwest-Europa	84	.77	.68	93	.74	.73
China	95	.81	.76	127	.63	.72
Oost-Europa	23	.75	.81	28	.69	.78
Zuid-Europa	124	.70	.80	147	.69	.74
Molukken	139	.73	.67	157	.68	.81
Antillen	58	.71	.72	74	.81	.84
Suriname: Hindoestanen	184	.76	.78	157	.68	.77
Suriname: Creolen	202	.68	.68	196	.71	.74
Turkije	431	.69	.71	534	.70	.76
Marokko	375	.69	.75	540	.68	.78

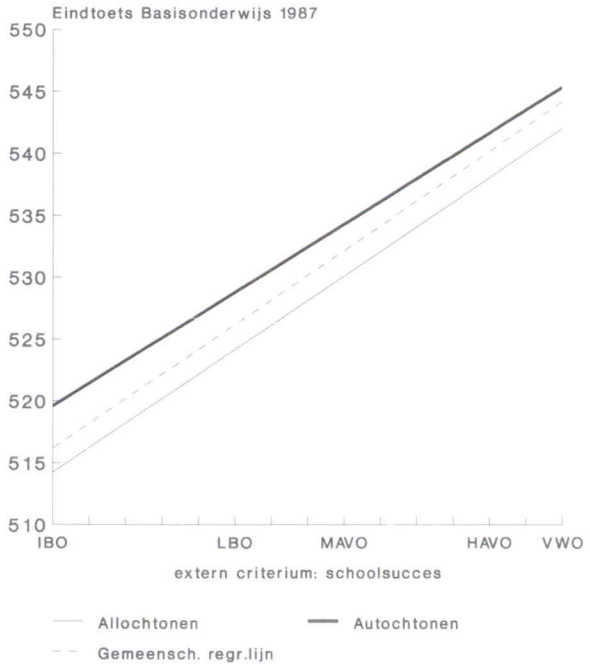
* *Alle correlaties zijn significant (p<.001)*

Uit tabel 5.3 blijkt dat de coëfficiënten per etnische groep van 1987 en 1989 soms aanzienlijk verschillen. Dit geeft aan dat we in het algemeen voorzichtig moeten zijn met het trekken van conclusies op basis van één meting. Gezien de coëfficiënten is de voorspellende waarde van advies basisschool en Cito-score bij allochtone leerlingen doorgaans lager dan bij autochtone. Verder blijkt dat in beide jaren zowel voor allochtone als voor autochtone leerlingen de voorspellende waarde van het advies basisschool hoger is dan van de Cito-score. Het advies basisschool verklaart vergeleken met de Cito-score in 1987 voor allochtone leerlingen 5% en in 1989 11% meer variantie in schoolsucces, voor de autochtone leerlingen is dat in beide jaren 11%.
Er kan niet gesteld worden dat de Eindtoets Basisonderwijs in vergelijking met het advies basisschool voor allochtone leerlingen minder geschikt is dan voor autochtone leerlingen. De gegevens uit 1987 laten zien dat de voorspellende waarde van de Eindtoets bij allochtone en autochtone leerlingen even hoog is. In 1989 is de voorspellende waarde van de Eindtoets bij allochtone leerlingen lager dan bij autochtone, maar hierin onderscheidt de toets zich niet van het advies basisschool.

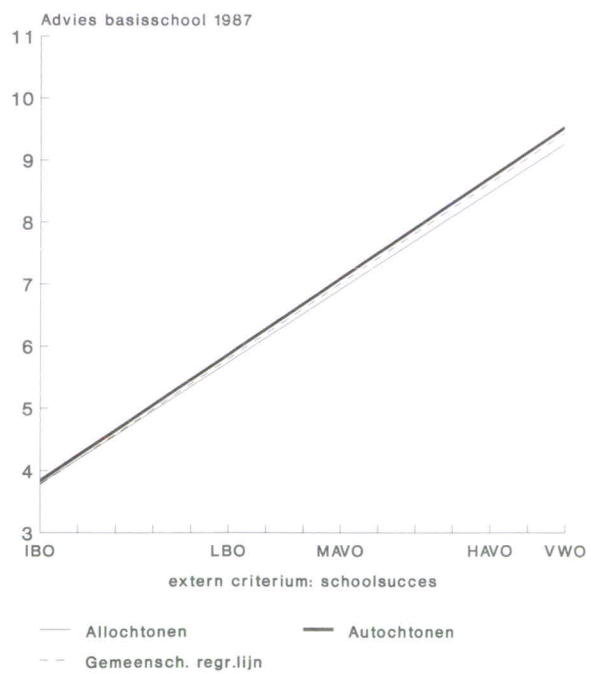
Om de predictieve validiteit van advies basisschool en de Eindtoets Basisonderwijs nauwkeuriger te kunnen onderzoeken is nagegaan hoe de regressielijnen van de beide onafhankelijke variabelen op de afhankelijke variabele schoolsucces lopen voor allochtone en autochtone leerlingen. In 1.2.1 is aangegeven dat de regressielijnen van voorspeller op extern criterium voor allochtone en autochtone leerlingen samenvallen wanneer de intercepten en de hellingen gelijk zijn. De regressielijnen vallen niet samen wanneer de intercepten (figuur 1.2) of de hellingen (figuur 1.3) verschillen of wanneer zowel de intercepten als de hellingen (figuur 1.4) verschillen (vgl. Cronbach, 1972; Jensen, 1980; Reynolds, 1982; Kok, 1988).
Van belang is of positie van allochtone dan wel van autochtone leerlingen op het extern criterium eventueel wordt over- of onderschat. Opgemerkt moet worden dat bij de Eindtoets Basisonderwijs voor de predictie van schoolsucces geen onderscheid gemaakt wordt tussen allochtone en autochtone leerlingen, er wordt voor alle leerlingen in feite de gemeenschappelijke regressievergelijking gehanteerd. In 2.1 is de verwachting uitgesproken dat het advies basisschool een overschatting geeft van het schoolsucces in het voortgezet onderwijs. Dit geeft aanleiding te veronderstellen dat de directeur basisschool voor allochtone en autochtone leerlingen in feite verschillende regressievergelijkingen hanteert.

De regressielijnen van het advies basisschool en de Eindtoets Basisonderwijs op schoolsucces voor zowel 1987 als 1989 zijn opgenomen in de figuren 5.2 – 5.5. De onafhankelijke variabele, Eindtoets Basisonderwijs of advies basisschool staat telkens op de Y-as, de afhankelijke variabele op de X-as. Als afhankelijke variabele wordt de schaal voor schoolsucces gehanteerd, die beschreven staat in 5.2.

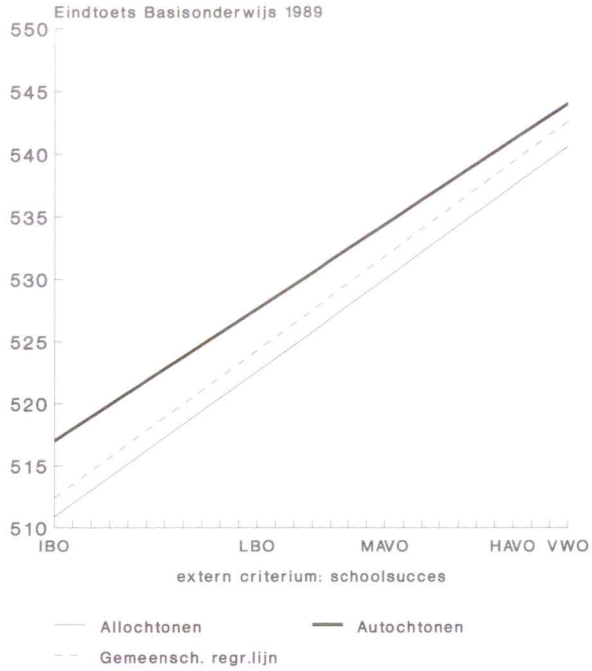
Figuur 5.2 Regressie van de Eindtoets 1987 op schoolsucces



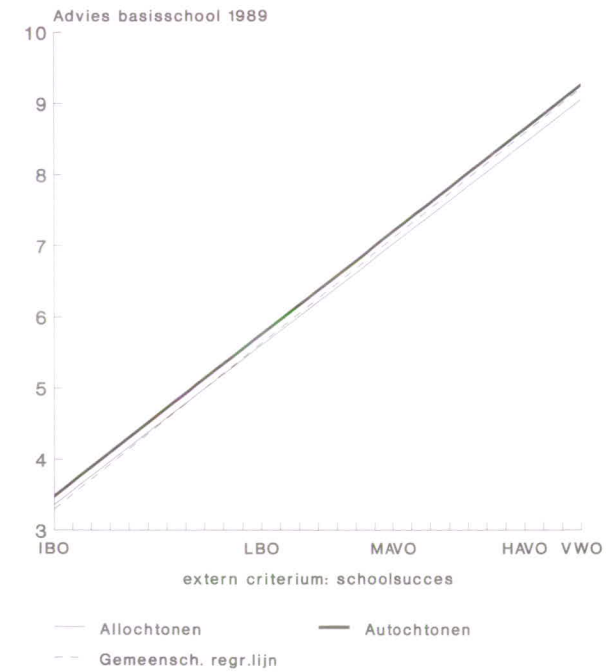
Figuur 5.3 Regressie van advies basisschool 1987 op schoolsucces



Figuur 5.4 Regressie van de Eindtoets 1989 op schoolsucces



Figuur 5.5 Regressie van advies basisschool 1989 op schoolsucces



Uit de figuren 5.2 – 5.5 blijkt dat de intercepten van allochtone en autochtone leerlingen bij de Eindtoets Basisonderwijs aanzienlijk meer verschillen dan die van het advies basisschool. De hellingen van beide groepen leerlingen ontlopen elkaar zowel bij de Eindtoets Basisonderwijs als bij het advies basisschool relatief weinig. De regressielijnen van de Eindtoets Basisonderwijs 1987 en 1989 op de schaal voor schoolsucces van allochtone en autochtone leerlingen verschillen significant ($p<.001$). Het verschil tussen de regressielijnen van het advies basisschool 1987 op de schaal voor schoolsucces van allochtone en autochtone leerlingen is net significant ($p<.05$), het verschil tussen de regressielijnen van het advies basisschool 1989 op de schaal voor schoolsucces is niet significant. Dit betekent dat bij het advies basisschool voor allochtone en autochtone leerlingen de gemeenschappelijke regressievergelijking – zeker in 1989 – niet geheel ten onrechte wordt gebruikt om de positie op het extern criterium te schatten. Ditzelfde geldt niet voor de Eindtoets Basisonderwijs. Als met de intercept en helling bij de gemiddelde Eindtoetsscore van de populatie nagegaan wordt welke positie op de schaal voor schoolsucces geschat wordt, dan blijkt dat bij dezelfde Eindtoetsscore de positie van allochtone leerlingen over 1987 en 1989 op de schaal voor schoolsucces gemiddeld 0.19 standaarddeviatie hoger ligt dan die van de autochtone leerlingen. De 0.19 standaarddeviatie op de schaal voor schoolsucces kan meer betekenis gegeven worden door deze afstand te relateren aan de afstand die onderwijstypen op deze schaal van elkaar vandaan liggen. Zo blijkt bijvoorbeeld dat de mate waarin de Eindtoets Basisonderwijs de positie van allochtone leerlingen in het voortgezet onderwijs overschat, te vergelijken is met 22% van de afstand die het onderwijstype MAVO 2e klas op deze schaal verwijderd is van HAVO 2e klas. Uit tabel 5.3 volgt dat correlatie tussen de Eindtoetsscore en de schaal voor schoolsucces lager ligt dan die van het advies basisschool. Uit de regressielijnen blijkt dat dit voor een deel te verklaren is uit het feit dat de Eindtoets Basisonderwijs het schoolsucces van allochtone leerlingen meer overschat dan het advies basisschool en dat het omgekeerde gebeurt bij de autochtone leerlingen. Nu moet reeds opgemerkt worden dat deze bevindingen moeilijk zijn te interpreteren, omdat zowel het advies basisschool als de Eindtoetsscore direct en indirect invloed uitoefenen op de positie die leerlingen na een jaar in het voortgezet onderwijs innemen. De directeur basisschool brengt immers een advies uit aan de ouders en aan de toelatingscommissie van het voortgezet onderwijs, de ouders en het kind beschikken over het advies basisschool en veelal over de Eindtoetsscore wanneer zij hun wens inzake de schoolkeuze definitief bepalen. De toelatingscommissie van het voortgezet onderwijs beoordeelt het advies basisschool, de Eindtoetsscore en de wens van ouders/kind alvorens een beslissing over de toelating te nemen. Over de interpretatie van de predictieve validiteit van advies basisschool en Eindtoetsscore in het kader van het schoolkeuzeproces wordt in hoofdstuk acht ingegaan.

5.4 De effecten van determinanten van schoolloopbanen van allochtone en autochtone leerlingen

De in 5.3 beschreven verbanden geven de relatie weer tussen een onafhankelijke en een afhankelijke variabele. Deze bivariate samenhangen zijn voor het bepalen van de effecten van determinanten van schoolloopbanen

minder informatief. Er kunnen zich situaties voordoen waarbij een hoge correlatie wordt gevonden tussen de onafhankelijke variabele A en de afhankelijke variabele X, maar uit nadere analyse zou kunnen blijken dat deze correlatie het gevolg is van het feit dat er een causaal verband bestaat met de onafhankelijke variabele B die aan variabele A voorafgaat en die zowel met variabele A als met de afhankelijke variabele X hoog correleert. De specifieke correlatie tussen variabele A en B en de afhankelijke variabele X kan de oorzaak zijn van de hoge correlatie tussen de variabelen A en X. De enkelvoudige regressie tussen de variabelen A en X kan in dit geval betrekking hebben op schijn-samenhang (vgl. Spaeth, 1975; Linn, 1983).
Voor het bepalen van de effecten van determinanten van schoolloopbanen is het van belang om de schijnverbanden op te sporen. Met multiple regressie-analyse kan nagegaan worden in welke mate de regressie tussen onafhankelijke variabele A en afhankelijke variabele X blijft bestaan, wanneer het effect van variabele B constant gehouden wordt. Het effect van variabele B wordt uitgepartialiseerd en de werkelijke regressie van A op X wordt geschat. Als er meer onafhankelijke variabelen zijn, moeten er meer regressie-analyses worden uitgevoerd. Het stelsel van regressie-analyses waarmee geprobeerd wordt een integraal beeld te geven van de effecten van de onafhankelijke op de afhankelijke variabelen, wordt een pad-analytisch model genoemd. Pad-analytische modellen worden ook gebruikt om de sterkte van de causale relaties in een model te schatten. De grootte van het effect van de ene variabele op de andere wordt uitgedrukt in de pad-coëfficiënten: dit zijn de gestandaardiseerde regressiecoëfficiënten (ß's of beta's). Binnen een causaal verklaringsmodel worden met pijlen de paden aangegeven waarlangs de effecten zich voltrekken. De effecten in het model kunnen direct en indirect zijn. Bij indirecte effecten loopt het effect van de ene onafhankelijke variabele via een andere onafhankelijke variabele naar de afhankelijke variabele. Het totale effect van een onafhankelijke op de afhankelijke variabele bestaat uit de som van de directe en indirecte effecten. De structuur in een causaal pad-analytisch model is recursief, de causale relaties zijn asymmetrisch: variabele D beïnvloedt wel variabele E, maar variabele E beïnvloedt niet D (Spaeth, 1975).

Pad-analyse is meer dan een statistische techniek. Pad-analyse veronderstelt een model dat gebaseerd is op een theorie. De theorie moet aangeven welke variabelen relevant zijn en tussen welke variabelen een causaal verband verondersteld mag worden. De theorie moet het principe aanreiken waarmee de variabelen geordend worden. Zo kan de volgorde in de tijd bepalen of er sprake kan zijn van causale relaties. De ervaring opgedaan in het basisonderwijs kan bijvoorbeeld wel de attitude in het voortgezet onderwijs beïnvloeden maar deze relatie kan niet recursief zijn.

5.4.1 Een schoolloopbaanmodel met het advies basisschool en de Cito-score

Om de determinanten van schoolsucces in het voortgezet onderwijs van allochtone en autochtone leerlingen te bepalen, zijn de onafhankelijke en afhankelijke variabelen in een model geplaatst. Het model is opgenomen in figuur 5.6. Als afhankelijke variabele geldt de schaal voor schoolsucces (zie 5.2). De onafhankelijke variabelen zijn in twee blokken geplaatst, blokken die geordend zijn volgens het principe van de volgorde in de tijd. De variabelen in

blok 1 kunnen het advies van de basisschool beïnvloeden en kunnen hun uitwerking hebben op de Cito-score, maar kunnen ook direct effect uitoefenen op de afhankelijke variabele schoolsucces. Het advies basisschool en de Cito-score (blok 2) hebben effect op het niveau dat de leerling aan het einde van het eerste leerjaar heeft bereikt. Er wordt een causale relatie verondersteld van de variabelen uit blok 1 met die uit blok 2 en alle variabelen kunnen direct of indirect een causale relatie met de afhankelijke variabele schoolsucces hebben.

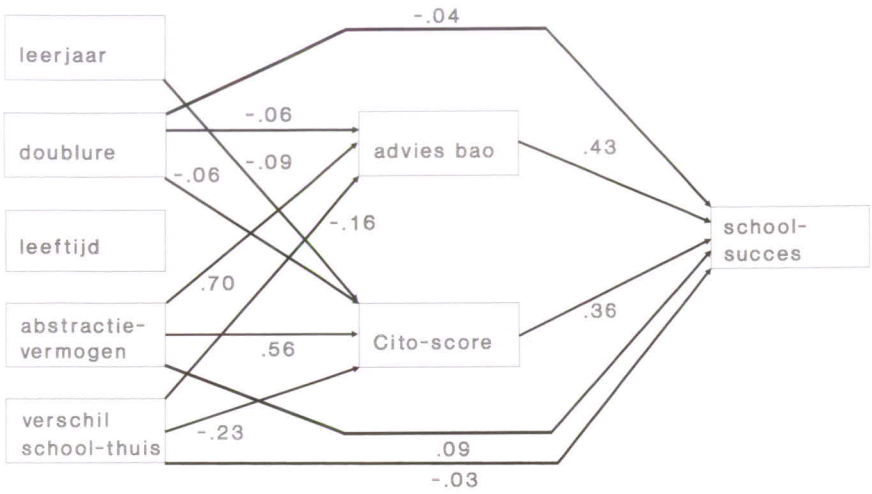
Figuur 5.6 Model met de relaties tussen de variabelen voor de pad-analyses

Onafhankelijke variabelen		Afhankelijke variabele
Blok 1	Blok 2	
startleerjaar	advies basisschool	schoolsucces
doublure	Cito-score	
leeftijd (alleen 1987)		
abstractievermogen		
verschil school-thuis		

De volgorde van de variabelen in het model heeft consequenties voor de volgorde waarin de multiple regressie-analyses worden uitgevoerd. Eerst is een analyse uitgevoerd met schoolsucces als afhankelijke variabele en alle andere variabelen als verklarende variabelen. Op deze wijze worden alle directe effecten op schoolsucces geschat. Daarna worden multiple regressie-analyses uitgevoerd om de directe effecten van de variabelen in blok 1 op het advies basisschool, respectievelijk de Cito-score (blok 2) te bepalen. Bij deze analyses fungeert dus het advies basisschool, respectievelijk de Cito-Eindtoetsscore als afhankelijke variabele. De analyses zijn voor allochtone en autochtone leerlingen afzonderlijk uitgevoerd. Opgemerkt moet worden dat de coëfficiënten bij de pijlen die vanuit 'verschil school-thuis' vertrekken in 1987 negatief zijn en in 1989 positief. Dit verschil heeft geen inhoudelijke betekenis, maar houdt verband met het feit dat de enquêtevraag in 1987 negatief is gesteld en in 1989 positief. De effecten van de onafhankelijke op de afhankelijke variabelen worden voor 1987 en 1989 in de figuren 5.7 tot en met 5.10 met pijlen weergegeven. In de figuren zijn alleen de significante ($p<.01$) effecten opgenomen. De vermelde verklaarde variantie heeft betrekking op de eerste analyse, waarbij schoolsucces de afhankelijke variabele is en alle andere variabelen als onafhankelijke fungeren. Dit model verklaart de meeste variantie. Het toevoegen van interactievariabelen (bijvoorbeeld: advies basisschool x Cito-score) heeft niet tot gevolg dat het percentage verklaarde variantie stijgt.

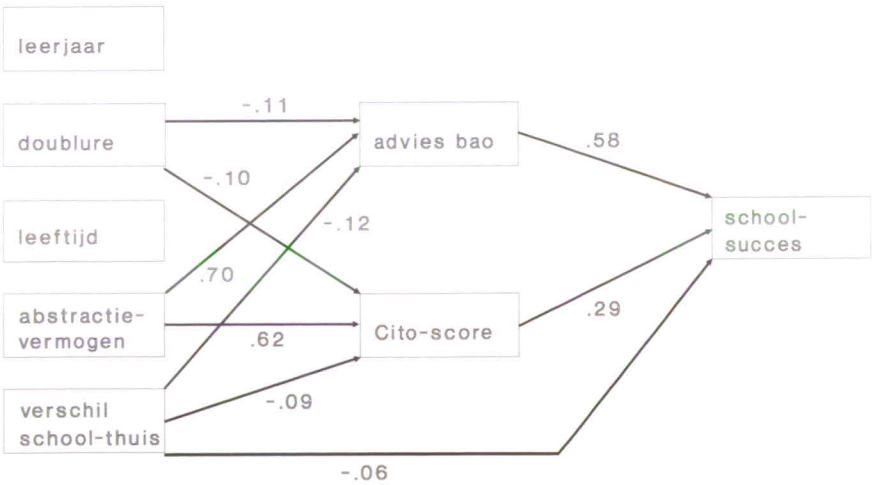
Figuur 5.7 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: allochtone leerlingen uit 1987

Verklaarde variantie (R^2) = .70



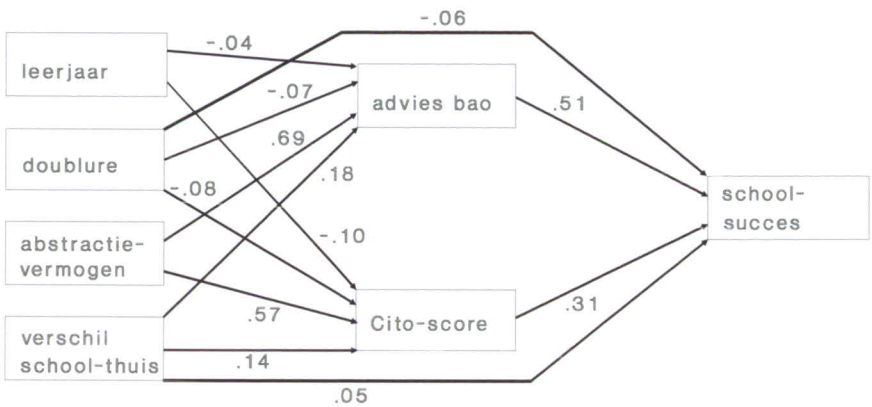
Figuur 5.8 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: autochtone leerlingen uit 1987

Verklaarde variantie (R^2) = .74



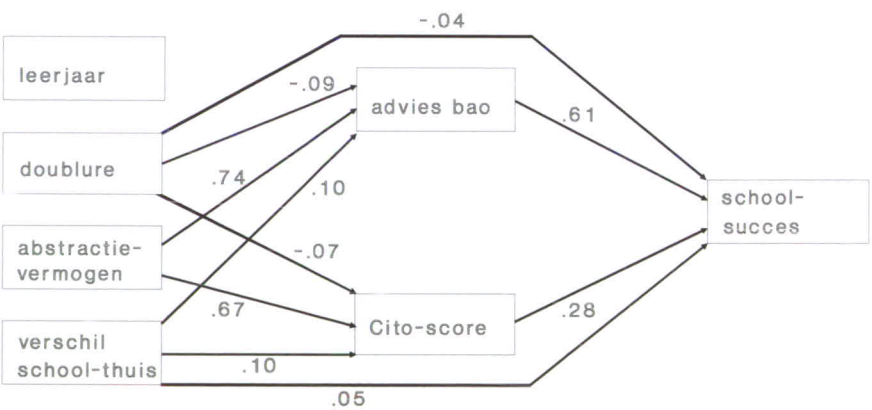
Figuur 5.9 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: allochtone leerlingen uit 1989

Verklaarde variantie (R^2) = .70



Figuur 5.10 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: autochtone leerlingen uit 1989

Verklaarde variantie (R^2) = .75



Uit de figuren 5.7 – 5.10 blijkt dat het schoolloopbaanmodel van autochtone leerlingen zowel in 1987 als in 1989 meer variantie in schoolsucces in het voortgezet onderwijs verklaart dan dat van allochtone leerlingen. Het effect van het advies basisschool op schoolsucces is in beide jaren zowel voor allochtone als autochtone leerlingen groter dan dat van de Cito-score. Verder blijkt uit de figuren 5.7 – 5.10 dat de Cito-score het schoolsucces van allochtone leerlingen beter voorspelt dan dat van autochtone leerlingen. De verschillen tussen de effecten van advies basisschool en Cito-score op schoolsucces zijn zowel in 1987 als in 1989 voor allochtone kleiner dan voor autochtone leerlingen. Het effect van het advies basisschool en dat van de Cito-score op schoolsucces verschilt in 1989 meer dan in 1987. Wanneer we in deze analyse de etnische groepen als dummy-variabele (vgl. Van de Vijver & Poortinga, 1991) invoeren, dan blijkt dat de verschillen tussen de effecten van advies en Cito-score op schoolsucces voor allochtone en autochtone leerlingen in beide jaren significant zijn ($p<.001$). Zoals in verband met zij-instromers te verwachten is, heeft het leerjaar waarin de leerling is gestart in het Nederlandse basisonderwijs, alleen een significant effect bij allochtone leerlingen. Het negatieve effect van de status van zij-instromer gaat voornamelijk richting Cito-score. Het negatieve (directe en indirecte) effect van doublure in het basisonderwijs op schoolsucces in het voortgezet onderwijs is in beide jaren voor allochtone en autochtone leerlingen vrijwel even groot.

Aan de hand van doublure in 1987 van allochtone leerlingen wordt nu toegelicht hoe het totale effect van doublure op schoolsucces bepaald wordt. Het directe effect van doublure op schoolsucces is in 1987 voor allochtone leerlingen -.04. Het indirecte effect van doublure op schoolsucces voor allochtone leerlingen via advies basisschool is (-.06 x .43 =) -.0258, het indirecte effect van doublure op schoolsucces voor allochtone leerlingen via Cito-score is (-.06 x .36 =) -.0216. Het totale effect van doublure op schoolsucces is derhalve -.09 (-.04 + -.0258 + -.0216). Het totale effect van doublure op schoolsucces in 1987 voor autochtone leerlingen is ook -.09. Het totale effect van doublure op schoolsucces in 1989 en van de overige onafhankelijke variabelen is opgenomen in tabel 5.4. Om de tabel te completeren zijn ook de directe effecten van advies basisonderwijs en Cito-score op schoolsucces vermeld.

Tabel 5.4 Totale effecten (β's) van de onafhankelijke variabelen op schoolsucces voor allochtone en autochtone leerlingen in 1987 en 1989 (n.s. = niet significant)

Onafhankelijke variabelen	1987		1989	
	Allocht.	Autocht.	Allocht.	Autocht.
leerjaar	-.03	n.s.	-.05	n.s.
doublure	-.09	-.09	-.12	-.11
leeftijd (alleen 1987)	n.s.	n.s.		
abstractievermogen	.59	.59	.53	.64
verschil school–thuis	-.18	-.16	.19	.14
advies basisschool	.43	.58	.51	.61
Cito-score	.36	.29	.31	.28

Uit tabel 5.4 blijkt dat het totale effect van 'abstractievermogen' op schoolsucces in beide jaren voor zowel allochtone als autochtone leerlingen bijzonder groot is. Uit de figuren 5.7 – 5.10 komt naar voren dat het effect van het 'abstractievermogen' zowel voor allochtone als autochtone leerlingen groter is op het advies basisschool dan op de Cito-score. Het effect van 'verschil school-thuis' op schoolsucces is in beide jaren voor allochtone leerlingen groter dan voor autochtone leerlingen. Het effect van 'verschil school-thuis' op advies basisschool, respectievelijk Cito-score is wisselend.

5.4.2 Een schoolloopbaanmodel met de toetsscores Taal, Rekenen en Informatieverwerking

In dit hoofdstuk is het begrip toetsscore tot nu toe beperkt tot de Cito-standaardscore, die het algemene prestatieniveau van de leerling aangeeft. Om na te gaan of de onderdelen Taal, Rekenen en Informatieverwerking van de Eindtoets Basisonderwijs 1987 en 1989 voor allochtone en autochtone leerlingen differentiële effecten hebben op schoolsucces in het voortgezet onderwijs, zijn hiermee ook pad-analyses uitgevoerd. Er zijn aanwijzingen dat de scores op Rekenen voor allochtone leerlingen een groter effect op schoolsucces hebben dan die voor Taal en Informatieverwerking. De aanwijzingen over de speciale positie van Rekenen zijn gebaseerd op de bevindingen van Kerkhoff (1988). Zij constateert dat de door de allochtone leerlingen (n=48) behaalde rapportcijfers voor rekenen een groter effect op het schoolkeuze-advies van de basisschool hebben dan de rapportcijfers voor taal. Bij dialectsprekende autochtone leerlingen (n=75) is het effect van de rapportcijfers voor taal op het advies van de basisschool het grootst, bij autochtone standaardtaalsprekende leerlingen (n=34) is het effect van de rapportcijfers voor taal en rekenen op het advies sterk vergelijkbaar. Hoewel het aantal waarnemingen in het onderzoek van Kerkhoff gering is, geven haar bevindingen aan dat leerkrachten bij het bepalen van het schoolkeuze-advies niet bij alle leerlingen de schoolvakken op dezelfde wijze laten meetellen. Opgemerkt moet worden dat in het onderhavige onderzoek niet het effect van schoolcijfers op het advies basisschool centraal staat, maar het effect van de drie toetsonderdelen op schoolsucces. Bovendien wordt geanalyseerd welke onafhankelijke variabelen effect hebben op de drie toetsonderdelen en direct en indirect op schoolsucces.

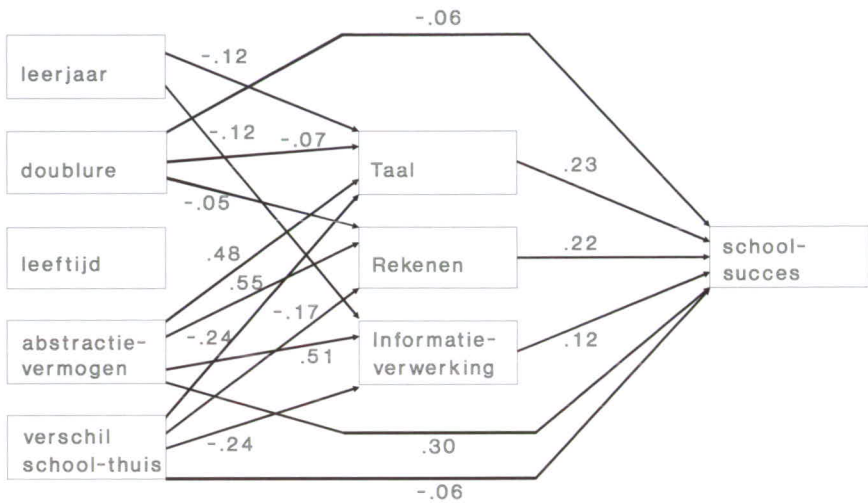
Figuur 5.11 Model met de relaties tussen de variabelen voor de pad-analyses

Onafhankelijke variabelen		Afhankelijke variabele
Blok 1	Blok 2	
startleerjaar	toetsscore Taal	schoolsucces
doublure	toetsscore Rekenen	
leeftijd (alleen 1987)	toetsscore Informatie-verwerking	
abstractievermogen		
verschil school-thuis		

Om de determinanten van schoolsucces te bepalen zijn net als in figuur 5.6 de onafhankelijke variabelen in twee blokken geplaatst, blokken die weer geordend zijn volgens het principe van de volgorde in de tijd. De variabelen in blok 1 kunnen de scores op Taal, Rekenen en Informatieverwerking beïnvloeden, maar kunnen ook direct effect uitoefenen op de afhankelijke variabele schoolsucces. De scores op de 3 toetsonderdelen (blok 2) kunnen voor allochtone en autochtone leerlingen differentiële effecten hebben op het niveau dat de leerlingen aan het einde van het eerste leerjaar bereiken. Er wordt een causale relatie verondersteld van de variabelen uit blok 1 op die uit blok 2 en alle variabelen kunnen direct of indirect een causale relatie op de afhankelijke variabele schoolsucces hebben. De multiple regressie-analyses voor het model in figuur 5.11 zijn in dezelfde volgorde uitgevoerd als die voor het model in figuur 5.6. Net als in 5.4.1 is om alle directe effecten op schoolsucces te schatten eerst een analyse uitgevoerd met schoolsucces als afhankelijke variabele en alle andere variabelen als verklarende variabelen. Daarna zijn multiple regressie-analyses uitgevoerd om de directe effecten van de variabelen in blok 1 op Taal, Rekenen en Informatieverwerking (blok 2) te schatten. Bij deze analyses fungeren de drie toetsonderdelen als afhankelijke variabelen. De significante ($p<.01$) effecten van de onafhankelijke op de afhankelijke variabelen worden voor 1987 en 1989 in de figuren 5.12 tot en met 5.15 weergegeven. De opzet van deze figuren is vergelijkbaar met die van 5.7 – 5.10 (zie 5.4.1). Ten aanzien van de figuren 5.12 – 5.15 moet opgemerkt worden dat het toevoegen van interactie-variabelen (bijvoorbeeld: Taal x Rekenen x Informatieverwerking) niet tot gevolg heeft dat het percentage verklaarde variantie meer dan 1% stijgt.

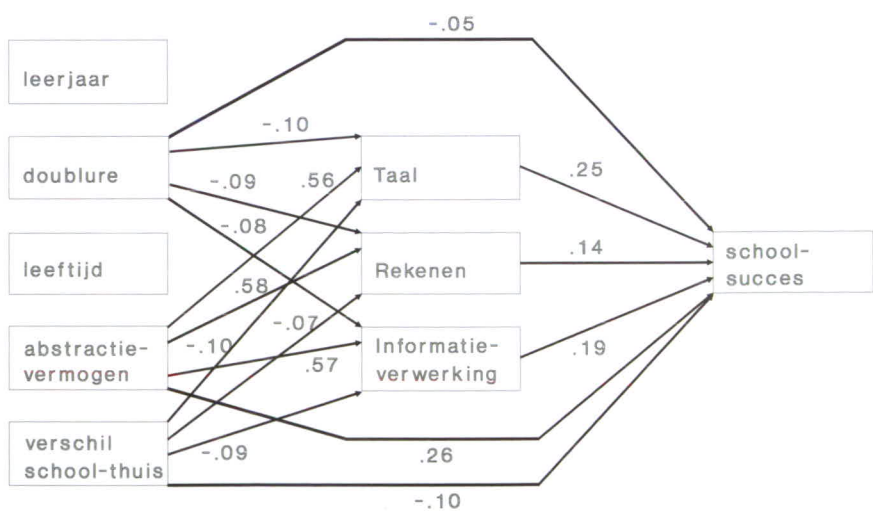
Figuur 5.12 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: allochtone leerlingen uit 1987

Verklaarde variantie (R^2) = .64



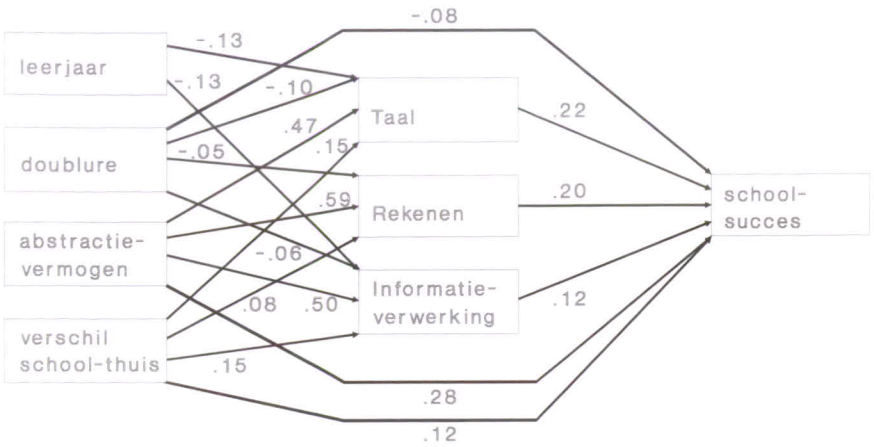
Figuur 5.13 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: autochtone leerlingen uit 1987

Verklaarde variantie (R^2) = .64



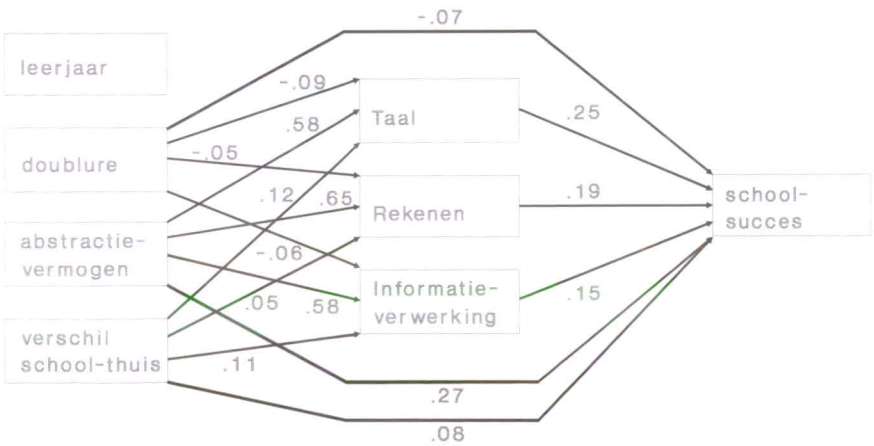
Figuur 5.14 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: allochtone leerlingen uit 1989

Verklaarde variantie (R^2) = .63



Figuur 5.15 Model met relaties tussen schoolsucces en de onafhankelijke variabelen: autochtone leerlingen uit 1989

Verklaarde variantie (R^2) = .66



De totale (directe en indirecte) effecten van de onafhankelijke variabelen uit blok 1 op schoolsucces zijn opgenomen in tabel 5.5. Om de tabel te completeren zijn ook de directe effecten van de toetsscores op Taal, Rekenen en Informatieverwerking op schoolsucces vermeld.

Tabel 5.5 Totale effecten (β's) van de onafhankelijke variabelen op schoolsucces voor allochtone en autochtone leerlingen in 1987 en 1989 (n.s. = niet significant)

Onafhankelijk variabelen	1987		1989	
	Allocht.	Autocht.	Allocht.	Autocht.
leerjaar	-.04	n.s.	-.04	n.s.
doublure	-.09	-.10	-.12	-.11
leeftijd (alleen 1987)	n.s.	n.s.		
abstractievermogen	.59	.59	.56	.63
verschil school-thuis	-.18	-.15	.19	.14
score Taal	.23	.25	.22	.25
score Rekenen	.22	.14	.20	.19
score Informatieverwerking	.12	.19	.12	.15

Uit de figuren 5.12 – 5.15 blijkt dat het effect van de taalscore op schoolsucces in vergelijking met de reken- en informatieverwerkingscore bij autochtone leerlingen groter is dan bij allochtone. Het effect van de taal- en rekenscore is bij allochtone leerlingen in beide jaren vrijwel gelijk. Het effect van de informatieverwerkingscore is zeker bij allochtone leerlingen gering. De rekenresultaten spelen bij de overgang naar het voortgezet onderwijs bij allochtone leerlingen een bijzondere rol. Kerkhoff (1988) vond dat bij allochtone leerlingen de rapportcijfers voor rekenen een groter effect hebben op het schoolkeuze-advies van de basisschool dan de rapportcijfers voor taal. Uit het onderhavige onderzoek blijkt dat de taal- en rekenscores van allochtone leerlingen een vrijwel gelijk effect hebben op schoolsucces, terwijl bij autochtone leerlingen het effect van de taalscore groter is. Deze resultaten vormen een aanvulling op de onderzoeksresultaten van Kerkhoff (1988), want de rapportcijfers van allochtone leerlingen voor het vak rekenen hebben niet alleen effect op het advies basisschool, ook de Citorekenscores hebben een aanzienlijk effect op schoolsucces.
Net als in de figuren 5.7 – 5.10 komt hier naar voren dat het leerjaar waarin de leerling in het Nederlandse basisonderwijs is gestart alleen een significant effect bij allochtone leerlingen heeft. Het 'verschil school-thuis' heeft bij allochtone leerlingen een groter effect op de scores voor Taal, Rekenen en Informatieverwerking dan bij autochtone leerlingen. Het totale effect van het 'verschil school-thuis' op schoolsucces is bij allochtone leerlingen ook groter. Doublure heeft zowel bij allochtone als bij autochtone leerlingen in 1989 een iets groter effect op schoolsucces dan in 1987 (vgl. tabellen 5.4 en 5.5). 'Abstractievermogen' heeft bij allochtone leerlingen een lager effect op de scores voor Taal, Rekenen en Informatieverwerking dan bij autochtone leerlingen. Het totale effect van 'abstractievermogen' op schoolsucces laat een wisselend beeld zien.

Wanneer we de verklaarde variantie van de schoolloopbanen in de figuren 5.7 – 5.10 vergelijken met die uit de figuren 5.12 – 5.15, dan blijkt dat de verklaarde variantie van de eerste modellen 6 – 10% hoger is. Dit verschil is te verklaren uit het feit, dat in de eerste modellen zowel het advies basisschool als

de Cito-(totaal)score in blok 2 zijn opgenomen en dat in de tweede modellen deze twee variabelen zijn vervangen door de scores op de drie Eindtoets-onderdelen, die elk deel uitmaken van de Cito-totaalscore.

5.4.3 Een schoolloopbaanmodel per onderscheiden etnische minderheidsgroep

Tot nu hebben we ons beperkt tot de subgroepen allochtone en autochtone leerlingen. Om een goed beeld te krijgen van de determinanten van schoolsucces van etnische minderheidsgroepen zijn ook per onderscheiden etnische groep multiple regressie-analyses uitgevoerd. Hierbij is eerst het model van figuur 5.6 gehanteerd, waarbij de directe effecten van alle onafhankelijke (blok 1 en blok 2) op de afhankelijke variabele schoolsucces zijn geschat. In tabel 5.6 staan per etnische groep de ß-coëfficiënten van de onafhankelijke variabelen met significante ($p<.01$) directe effecten op schoolsucces en het percentage variantie dat de onafhankelijke variabelen samen in schoolsucces verklaren.

Tabel 5.6 Per etnische groep de directe significante ($p<.01$) effecten (ß's) van de onafhankelijke variabelen op schoolsucces in 1987 resp. 1989

Groep	n	R^2	dou-blure	abstr. verm	versch. school-thuis	advies bao	Cito-score
1987							
Autochtonen	3274	.74			-.06	.58	.29
Noordwest-Europa	84	.74			-.25	.37	.49
China	95	.72					.51
Oost-Europa	23	.79					
Zuid-Europa	124	.67				.61	
Molukken	139	.62				.33	.55
Antillen	58	.69				.38	.36
Suriname: Hindoest.	184	.71				.40	.41
Suriname: Creolen	202	.61	-.15		-.12	.35	.37
Turkije	431	.62			-.08	.38	.37
Marokko	375	.65		.13		.45	.30
1989							
Autochtonen	3405	.75	-.04		.05	.61	.28
Noordwest-Europa	93	.66				.51	.48
China	127	.60				.53	.34
Oost-Europa	28	.66					
Zuid-Europa	147	.63				.46	.35
Molukken	157	.70	-.13			.57	
Antillen	74	.76				.61	
Suriname: Hindoest.	157	.63				.60	.25
Suriname: Creolen	196	.66			.13	.39	.38
Turkije	534	.64				.43	.30
Marokko	540	.66	-.09			.59	.30

Uit tabel 5.6 blijkt dat het percentage verklaarde variantie (R^2) tussen de onderscheiden etnische minderheidsgroepen varieert van 60 tot 79%. De verschillen tussen het percentage verklaarde variantie in 1987 en 1989 varieert per etnische groep van 1% (autochtonen en Marokko) tot 13% (Oost-Europa). De effecten van de onafhankelijke variabelen op schoolsucces in 1987 en 1989 laten in hun onderlinge verhouding niet altijd een consistent beeld zien. Bij autochtone leerlingen en bij leerlingen met ouders uit Zuid-Europa en Marokko blijkt dat in beide jaren het effect van het advies basisschool dominant is. Bij de Surinaams-Creoolse leerlingen is het effect van het advies basisschool en de Cito-score in beide jaren vrijwel even groot. Verder valt in tabel 5.6 op dat bij autochtone en bij Surinaams-Creoolse leerlingen het 'verschil school-thuis' in beide jaren een – gering – effect heeft op schoolsucces. Bij de overige onderscheiden etnische minderheidsgroepen zijn de effecten van de onafhankelijke variabelen in 1987 en 1989 te wisselend om de verhouding tussen de determinanten van schoolsucces aan te geven.
De effecten in tabel 5.6 hebben betrekking op de significante ($p<.01$) directe effecten op schoolsucces. Opgemerkt wordt dat de totale effecten (directe + indirecte) van de onafhankelijke variabelen op schoolsucces hiervan kunnen afwijken.

Om na te gaan of onderdelen Taal, Rekenen en Informatieverwerking van de Eindtoets Basisonderwijs 1987 en 1989 voor de onderscheiden etnische groepen differentiële effecten hebben op schoolsucces in het voortgezet onderwijs, zijn hiermee ook pad-analyses uitgevoerd. Hierbij is het model van figuur 5.11 gehanteerd, waarbij de directe effecten van alle onafhankelijke (blok 1 en blok 2) op de afhankelijke variabele schoolsucces zijn geschat. In tabel 5.7 staan per etnische groep de ß-coëfficiënten van de onafhankelijke variabelen met significante ($p<.01$) directe effecten op schoolsucces en het percentage variantie dat de onafhankelijke variabelen samen in schoolsucces verklaren.

Tabel 5.7 Per etnische groep de directe significante (p<.01) effecten (β's) van de onafhankelijke variabelen op schoolsucces in 1987 resp. 1989

Groep	R^2	dou-blure	abstr. verm	versch. school-thuis	Taal	Reke-nen	Infor-matie-verw.
1987							
Autochtonen	.64	-.05	.26	-.10	.25	.14	.19
Noordwest-Europa	.74		.25	-.25		.43	
China	.69		.24			.25	
Oost-Europa	.81						
Zuid-Europa	.59		.36		.37		
Molukken	.60		.		34	.28	
Antillen	.64						
Suriname: Hindoest.	.66		.28	-.14	.23	.24	
Suriname: Creolen	.56	-.17	.25				
Turkije	.57		.27	-.10	.17	.16	.22
Marokko	.57		.35		.16	.18	.17
1989							
Autochtonen	.66	-.07	.27	.08	.25	.19	.15
Noordwest-Europa	.61				.49		
China	.49			.21		.32	
Oost-Europa	.76				.95		
Zuid-Europa	.57				.31	.29	
Molukken	.62		.35				
Antillen	.71						
Suriname: Hindoest.	.55		.29		.24		
Suriname: Creolen	.61		.27	.15	.24		
Turkije	.58		.33	.11	.18	.15	.15
Marokko	.55	-.11	.25	.09	.15	.23	.16

Uit tabel 5.7 blijkt dat het percentage verklaarde variantie (R^2) varieert van 49 – 81%. De verschillen tussen het percentage verklaarde variantie in beide jaren varieert per etnische groep van 1% (Turkije) – 20% (China). Verder volgt uit tabel 5.7 dat het effect van de informatieverwerkingscore op schoolsucces bij de meeste etnische groepen niet significant is; bij autochtone leerlingen en bij de leerlingen uit Turkije en Marokko is dit wel het geval. Ten aanzien van de Chinese leerlingen is op te merken dat alleen de rekenscores effect op schoolsucces hebben, het effect van 'abstractievermogen' en 'verschil school-thuis' is wisselend. Bij de Turkse en Marokkaanse leerlingen zijn de effecten van de drie toetsonderdelen op schoolsucces in beide jaren vergelijkbaar, terwijl ook 'abstractievermogen' en 'verschil school-thuis' doorgaans een significant effect op schoolsucces hebben.

5.5 Samenvatting

In Hoofdstuk 1 is aangegeven dat onderzoek naar toetsbias opgevat is als het nagaan van de predictieve validiteit van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen in vergelijking met die van het advies van de basisschool. In 5.1 wordt na vergelijking van de produkt-moment correlatie-coëfficiënten en de η-coëfficiënten (eta-coëfficiënten) van de onafhankelijke variabelen geconcludeerd dat er geen betekenisvolle schattingsfouten worden gemaakt wanneer voor de meting van de onafhankelijke variabelen het intervalniveau wordt aanvaard. In 5.2 wordt de constructie van een schaal voor schoolsucces, die in het onderzoek als afhankelijke variabele functioneert, verantwoord. De schaal voor schoolsucces wordt gedefinieerd door aan de onderwijsposities die de leerlingen in het voortgezet onderwijs innemen de waarde toe te kennen van de gemiddelde Cito-standaardscore van de leerlingen in dat onderwijstype.

In dit hoofdstuk komt de derde onderzoeksvraag aan bod.

3 Hoe hoog is voor allochtone en autochtone leerlingen de voorspellende waarde van de Eindtoets Basisonderwijs in vergelijking met het advies van de basisschool.

Voor de beantwoording van deze vraag zijn eerst de produkt-moment correlatie-coëfficiënten tussen deze twee variabelen en de schaal voor schoolsucces berekend (5.3). Hieruit blijkt dat de voorspellende waarde van deze variabelen bij allochtone leerlingen doorgaans lager is dan bij autochtone. Het voorspellen van schoolsucces voor allochtone leerlingen gaat dus minder trefzeker dan voor autochtone. Het advies basisschool verklaart meer variantie in schoolsucces dan de Cito-score. Uit de analyses volgt dat de Eindtoets Basisonderwijs voor allochtone leerlingen niet minder bruikbaar is dan voor autochtone. De Eindtoets voorspelt in 1987 het schoolsucces van allochtone en autochtone leerlingen even goed. In 1989 is de voorspellende waarde van de toets voor allochtone leerlingen iets lager dan voor autochtone leerlingen, maar hierin verschilt de toets niet met het advies basisschool.
Om de predictieve validiteit van het advies basisschool en Eindtoets Basisonderwijs nader te analyseren is nagegaan hoe de regressielijnen van de beide onafhankelijke variabelen op de afhankelijke variabele schoolsucces lopen voor allochtone en autochtone leerlingen. Uit deze analyses blijkt dat de regressielijnen van allochtone en autochtone leerlingen van de Eindtoets Basisonderwijs op de schaal voor schoolsucces significant verschillen ($p<.001$). Het verschil tussen de regressielijnen van allochtone en autochtone leerlingen van het advies basisschool op schoolsucces is in 1987 net significant ($p<.05$), dit verschil is in 1989 niet significant. De Eindtoets Basisonderwijs blijkt het schoolsucces van allochtone leerlingen te overschatten, dat van autochtone leerlingen te onderschatten. De positie van allochtone leerlingen op de schaal voor schoolsucces schat de Eindtoets Basisonderwijs gemiddeld 0.19 standaard-deviatie hoger dan die van autochtone leerlingen. De lagere voorspellende waarde van de toets in vergelijking met het advies basisschool is voor een deel te verklaren uit het feit dat de Eindtoets Basisonderwijs het schoolsucces van allochtone leerlingen meer overschat en dat van autochtone leerlingen meer

onderschat dan het advies basisonderwijs. Deze onderzoeksresultaten zijn moeilijk te interpreteren, omdat zowel het advies basisschool als de Eindtoets Basisonderwijs effect uitoefenen op de positie die beide groepen leerlingen na een jaar in het voortgezet onderwijs innemen. In hoofdstuk acht wordt nader ingegaan op de interpretatie van de verschillen in predictieve validiteit.

Om inzicht te krijgen in het causale effect van de onafhankelijke variabelen op de afhankelijke variabele schoolsucces zijn pad-analyses uitgevoerd (5.4). Hiervoor is een schoolloopbaanmodel opgesteld waarin de onafhankelijke variabelen verdeeld zijn over twee blokken, die geordend zijn volgens volgorde in de tijd. De variabelen in blok 1 kunnen effect hebben op de variabelen in blok 2 en op de afhankelijke variabele schoolsucces. De variabelen in blok 2 kunnen effect hebben op schoolsucces.
Uit de pad-analyses blijkt dat de schoolloopbaanmodellen zowel in 1987 als in 1989 meer variantie in schoolsucces verklaren bij autochtone dan bij allochtone leerlingen. Hieruit volgt dat het schoolsucces van allochtone leerlingen minder trefzeker is te voorspellen dan dat van autochtone leerlingen. Het effect van het advies basisschool op schoolsucces is in beide jaren groter dan het effect van de Cito-score. Verder blijkt dat de Cito-score het schoolsucces van allochtone leerlingen beter voorspelt dan dat van autochtone leerlingen. Het leerjaar waarin de leerling is gestart in het Nederlandse basisonderwijs, heeft, zoals verwacht, alleen een significant effect op schoolsucces bij allochtone leerlingen. Het effect van het 'abstractievermogen' op schoolsucces is groot, zowel voor allochtone als voor autochtone leerlingen. Het effect van het 'verschil school-thuis' op schoolsucces is voor allochtone leerlingen groter dan voor autochtone leerlingen.
Om na te gaan of de toetsonderdelen Taal, Rekenen en Informatieverwerking bij allochtone en autochtone leerlingen differentiële effecten hebben op schoolsucces, zijn ook pad-analyses met de drie toetsonderdelen uitgevoerd. Hieruit blijkt dat het effect van de taalscore op schoolsucces bij autochtone leerlingen groter is dan bij allochtone, bij allochtone leerlingen is het effect van de taal- en rekenscore vrijwel gelijk. Het effect van de informatieverwerking-score is bij allochtone leerlingen geringer.
Om een goed beeld te krijgen van de determinanten van schoolsucces van de onderscheiden etnische minderheidsgroepen zijn ook pad-analyses per etnische groep uitgevoerd. De effecten van de verschillende onafhankelijke variabelen op schoolsucces laten in hun onderlinge verhouding voor de onderscheiden etnische groepen over het algemeen geen consistent beeld zien. Bij autochtone leerlingen en bij leerlingen met ouders uit Zuid-Europa en Marokko blijkt dat zowel in 1987 als in 1989 het effect van het advies basisschool dominant is. Bij de Surinaams-Creoolse leerlingen is het effect van het advies basisschool en de Cito-score in beide jaren vrijwel even groot. Verder valt op dat het effect van de informatieverwerkingscore op schoolsucces alleen bij autochtone, Turkse en Marokkaanse leerlingen significant is. Bij de Chinese leerlingen heeft in 1987 en 1989 alleen de rekenscore effect op schoolsucces. De Chinese leerlingen behalen de hoogste rekenscores (zie 4.2) en deze scores overtreffen in beide jaren de taal- en informatieverwerkingscores met betrekking tot het effect op schoolsucces.

6 Itembias in de Eindtoets Basisonderwijs 1987 en 1989

In empirisch onderzoek is het gebruikelijk om schoolvorderingentoetsen te gebruiken om de vaardigheid van leerlingen in bepaalde vormingsgebieden (meestal taal en rekenen) te meten. Over het algemeen wordt aangenomen dat de gehanteerde schoolvorderingentoetsen een geschikt middel zijn om de bedoelde vaardigheid van zowel allochtone als autochtone leerlingen te meten. De verschillen tussen de scores van de onderscheiden groepen kunnen toegeschreven worden aan verschillen in de te meten vaardigheid, maar ze kunnen ook een artefact zijn van de wijze waarop die vaardigheid is gemeten. Wanneer blijkt dat toetsopgaven over bijvoorbeeld tekstbegrip voor allochtone leerlingen moeilijker zijn dan voor autochtone, dan is het mogelijk dat de toets zijn functie naar behoren vervult: het maken van onderscheid tussen goede en minder goede leerlingen op het gebied van tekstbegrip. Voor het juist beantwoorden van die tekstbegrip-opgaven kunnen echter nog andere vaardigheden nodig zijn dan de vaardigheid die de items beogen te meten. Zo is er met betrekking tot begrijpend lezen een bepaalde voorkennis nodig over hetgeen in de tekst aan de orde wordt gesteld. Wanneer de benodigde voorkennis of additionele vaardigheid niet bij alle onderscheiden groepen in gelijke mate aanwezig zijn, kunnen we zeggen dat de toets niet voor alle leerlingen constructvalide is. De kans om de toetsopgave in kwestie juist te beantwoorden, is dan niet gelijk voor leerlingen die weliswaar even vaardig zijn in het beantwoorden van opgaven over tekstbegrip, maar verschillen met betrekking tot de aard en het niveau van de benodigde additionele vaardigheden. Het item meet dan geen ééndimensionele vaardigheid (vgl. 2.2). Er is sprake van itembias wanneer leerlingen uit onderscheiden subgroepen maar met dezelfde vaardigheid een ongelijke kans hebben om het betreffende item goed te beantwoorden. Wanneer leerlingen uit verschillende etnische groepen aan de hand van een relevant criterium ingedeeld zijn in groepen met een gelijk vaardigheidsniveau, kan nagegaan worden of de kans op een goed antwoord voor allochtone en autochtone leerlingen gelijk is. Er moet dus een criterium beschikbaar zijn om leerlingen te rangschikken in niveaugroepen, dat hetzelfde construct meet als de op itembias te onderzoeken items beogen te meten. Vervolgens kan met bepaalde statistische procedures vastgesteld worden welke items partijdig zijn voor een bepaalde etnische groep (vgl. 1.2.2).

Kok (1988) onderscheidt in onderzoek naar itembias twee fasen:
de detectie- en de verklaringsfase. In de detectiefase wordt met een statistische procedure bepaald welke items partijdig zijn. Daarna kan in de verklaringsfase via inhoudsanalyse nagegaan worden welke elementen uit een item naar alle waarschijnlijkheid de itembias veroorzaken. In het onderhavige onderzoek zijn beide complementaire fasen doorlopen.

In 6.1 worden de statistische procedures verantwoord die gehanteerd zijn om te bepalen welke items van de Eindtoets Basisonderwijs 1987 en 1989 partijdig zijn voor leerlingen uit etnische minderheidsgroepen. In 6.2 worden de resultaten van de analyses naar itembias besproken. Dit hoofdstuk wordt afgesloten met

een samenvatting en conclusie (6.3). In hoofdstuk zeven wordt ingegaan op de inhoudelijke analyse van partijdige items met het doel bronnen van itembias op te sporen.

6.1 De itembiasdetectieprocedure

Uit de definitie van itembias volgt dat met betrekking tot een bepaald item vastgesteld moet worden of leerlingen uit onderscheiden subgroepen, maar met hetzelfde vaardigheidsniveau, een gelijke kans hebben om het item goed te beantwoorden.
Om de Eindtoets Basisonderwijs 1987 en 1989 op itembias te onderzoeken, moet een statistische procedure gekozen worden. In 1.2.2 is aangegeven dat de itembiasdetectieprocedures in twee groepen verdeeld kunnen worden: procedures gebaseerd op de itemresponsetheorie (IRT) of op de klassieke testtheorie. Voor de keuze en voor het hanteren van de itembiasdetectie-procedure is het van belang om na te gaan volgens welke theorie de te onderzoeken toets wordt samengesteld en (psychometrisch) geanalyseerd. De constructie en de psychometrische analyse van de Eindtoets Basisonderwijs is gebaseerd op de klassieke testtheorie (zie 3.1). De verdeling van het totaal aantal items over de drie toetsonderdelen (60 opgaven Taal, 60 opgaven Rekenen en 60 opgaven Informatieverwerking) gebeurt op vakinhoudelijke gronden. Bij elk toetsonderdeel wordt ervan uitgegaan dat het totaal aantal goed gemaakte opgaven een goede schatting is van de te meten vaardigheid. Het ligt voor de hand om voor het onderzoek naar itembias in de Eindtoets Basisonderwijs een itembiasdetectieprocedure te hanteren die gebaseerd is op de klassieke testtheorie.
Het is ook mogelijk om voor de Eindtoets Basisonderwijs een op een IRT-model gebaseerde procedure te kiezen, maar dan zal eerst vastgesteld moeten worden welke items een eendimensionele schaal vormen. Pas dan kan per schaal bepaald worden welke items partijdig zijn voor leerlingen uit etnische minderheidsgroepen.
Omdat gebleken is dat IRT- en klassieke testtheorieprocedures niet tot identieke resultaten leiden (Skaggs & Lissitz, 1988; Kok, 1988; Hambleton & Rogers, 1989; Hills, 1989; Camilli & Smith, 1990; Bügel & Glas, 1991), worden in het onderhavige onderzoek beide procedures gehanteerd.
In 6.1.1 wordt ingegaan op itembiasdetectie op basis van de klassieke testtheorie, in 6.1.2 volgt itembiasdetectie volgens het IRT-model. In 6.1.3 wordt de opzet van de uitgevoerde analyses naar itembias aan de orde gesteld.

6.1.1 Klassieke testtheorieprocedures

Bij de itembiasdetectieprocedures die gebaseerd zijn op de klassieke testtheorie wordt van de aanname uitgegaan dat de totaalscore een adequate schatting is van de te meten vaardigheid. In de Verenigde Staten wordt als zodanig de Mantel-Haenszel-procedure veelvuldig toegepast (Holland & Thayer, 1986; Skaggs & Lissitz, 1988; Tatsuoka e.a., 1988; Hills, 1989; Hambleton & Rogers, 1989; Raju e.a., 1989; Zwick, 1990; Camilli & Smith, 1990; Engelhard e.a., 1990; Scheuneman & Gerritz, 1990; Clauser e.a., 1991; Dorans & Holland, 1992; Ackerman & Evans, 1992; Holland & Wainer, 1993). Maar ook in ons land

krijgt deze procedure aandacht (Kok, 1988; Verhelst, 1988; Uiterwijk, 1990a; Bügel & Glas; 1991; Van de Vijver, 1991; Glas & Ouborg, 1993). De Mantel-Haenszel-procedure toetst de hypothese dat de moeilijkheidsgraad van een item bij twee groepen met dezelfde gemiddelde vaardigheid gelijk is. Leerlingen uit twee subgroepen (bijvoorbeeld: allochtone en autochtone leerlingen), maar met eenzelfde totaalscore, worden in niveaugroepen ingedeeld. Per item wordt voor elke niveaugroep een 2x2-tabel opgesteld.

Figuur 6.1 Vierveldentabel van niveaugroep j op item i

	goed	fout	
Referentiegroep (autochtonen)	A	B	T_r
Onderzoeksgroep (allochtonen)	C	D	T_o
	m_g	m_f	T_j

In niveaugroep j zitten T_j leerlingen, waarvan T_r de referentiegroep (autochtonen) vormt en T_o de onderzoeksgroep (allochtonen). Van de autochtone leerlingen hebben A-leerlingen het juiste antwoord gekozen, terwijl B-leerlingen kozen voor het foute antwoord. Bij de allochtone leerlingen zijn dit de C-, respectievelijk de D-leerlingen. In totaal zijn er bij item i m_g goede antwoorden en m_f foute antwoorden gegeven. Indien item i niet partijdig is, dan is te verwachten dat $A/T_r \approx C/T_o \approx m_g/T_j$.
Indien item i makkelijker is in de referentiegroep dan in de onderzoeksgroep, dan zullen A en D relatief groot zijn en wordt verwacht dat

$$\hat{\alpha} = \frac{AD}{CB} > 1 \quad \text{of} \quad \log \hat{\alpha} > 0.$$

De grootheid $\hat{\alpha}$ wordt de 'odds-ratio' (ratio van kansen) genoemd en de logaritme de 'log-odds-ratio'. Indien het item onpartijdig is dan verwachten we dat

$$\hat{\alpha} \approx 1 \quad \text{of} \quad \log \hat{\alpha} \approx 0.$$

De resultaten van de verschillende niveaugroepen worden samengevoegd en de toetsingsgrootheid is dan

$$\hat{\alpha}_{MH} = \frac{\Sigma\, AD/T_j}{\Sigma\, CB/T_j}.$$

Indien er geen itembias voorkomt en dus $\hat{\alpha}_{MH} = 1$, is $\log \hat{\alpha}_{MH}$ normaal verdeeld met een gemiddelde van 0 en een standaarddeviatie $SE(\log \hat{\alpha}_{MH})$, zodat de gestandaardiseerde log-odds-ratio $z = \log (\hat{\alpha}_{MH})/SE(\log \hat{\alpha}_{MH})$ bij benadering

standaardnormaal verdeeld is (Holland & Thayer, 1986; Verhelst, 1988; Bügel & Glas, 1991; Glas & Ouborg, 1993).
Bij een significantieniveau van 1%, wordt
- bij $z > 2.58$ besloten dat item i moeilijker is voor de onderzoeksgroep;
- bij $z < -2.58$ besloten dat het item moeilijker is voor de referentiegroep.

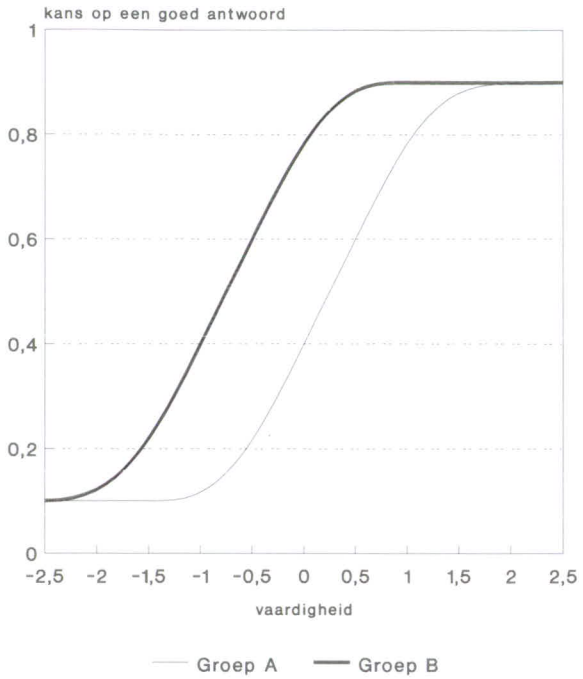
Bij de Mantel-Haenszel-procedure worden niveaugroepen gevormd aan de hand van de totaalscore. De aanwezigheid van partijdige items maakt de totaalscore minder geschikt als indicator van de vaardigheid van de leerlingen. Het is mogelijk door middel van een iteratief proces de totaalscore te 'zuiveren' van partijdige items (zie 6.1.3).

6.1.2 Itemresponsetheorie

De itemresponsetheorie (IRT) gaat er van uit dat de geobserveerde itemresponsen van een bepaalde populatie verklaard kunnen worden vanuit één onderliggende vaardigheid. Als het IRT-model past, wordt de kans op een goed antwoord beschreven als een functie van persoons- en itemparameters. Leerlingen met dezelfde score op de onderliggende vaardigheid of de latente trek hebben dan een gelijke kans om het item goed te beantwoorden, onafhankelijk van de populatie waartoe ze behoren. Als het IRT-model past, kunnen de parameters van de itemkarakteristieke curve (item characteristic curve of ICC) geschat worden. De ICC representeert de kans het item goed te beantwoorden als een functie van de toename in vaardigheid (zie 1.2.2). De ICC wordt veelal door drie parameters beschreven:
- de moeilijkheidsgraadparameter, die het vaardigheidsniveau aangeeft;
- de discriminatieparameter, die aangeeft in welke mate de kans op een goed antwoord stijgt, naarmate de vaardigheid toeneemt;
- de raadparameter, die de kans weergeeft dat de toetsdeelnemer het item goed beantwoordt door te raden.

De afstand tussen de ICC's van twee subgroepen geeft aan in welke mate de kans op een goed antwoord op elk vaardigheidsniveau verschilt. In figuur 6.2 worden als voorbeeld de ICC's van twee groepen weergegeven. Een item is onder een IRT-model partijdig wanneer de verschillen tussen de parameters van de ICC's van twee subgroepen significant zijn (Skaggs & Lissitz, 1988; Hambleton & Rogers, 1989; Hills, 1989; Mellenbergh, 1989; Camilli & Smith, 1990; Bügel & Glas, 1991).

Figuur 6.2 De 'item characteristic curves' van twee groepen



Mellenbergh (1989) noemt een aantal aandachtspunten die bij het gebruik van een IRT-model als itembiasdetectieprocedure van belang zijn. Deze aandachtspunten worden hieronder besproken.

Ten eerste moet er een bepaald IRT-model voor het onderzoek naar itembias gekozen worden. De meest gebruikelijke zijn de een- (het zgn. Raschmodel), twee- en drie-parametermodellen. In het Rasch-model is het aantal goed beantwoorde items een voldoende toetsingsgrootheid om de latente trek van een persoon te schatten. Bij het tweeparameter-model is tevens de discriminatieparameter opgenomen en in het drieparameter-model is ook de raadkans verdisconteerd.

Ten tweede wordt aangegeven dat de stabiliteit van de toetsingsgrootheid van de itembiasdetectieprocedure niet altijd perfect is. Deze stabiliteit kan in beeld gebracht worden door bijvoorbeeld uit elke subgroep twee steekproeven te trekken en de indices van de items in de beide steekproeven te vergelijken. Mellenbergh (1989) en Bügel & Glas (1991) geven hiervoor de volgende procedure.
Eerst worden uit elke subgroep twee steekproeven getrokken (bijvoorbeeld autochtonen 1 en 2; allochtonen 1 en 2). Vervolgens wordt voor de beide steekproeven van een bepaalde subgroep (autochtonen 1 en 2) onderzocht welke eendimensionele schalen in de items te onderscheiden zijn. De schalen die voor beide steekproeven autochtone leerlingen passen vormen de 'baseline' voor het onderzoek naar itembias (zie Mellenbergh, 1989). In de tweede stap wordt nagegaan of de gevonden schalen ook van toepassing zijn voor de twee

steekproeven uit de tweede subgroep: allochtonen 1 en 2. Hiertoe worden de parameters van de items van de schaal geschat op de itemresponsen van beide subgroepen (allochtonen 1 en autochtonen 1; allochtonen 2 en autochtonen 2). Bügel & Glas (1991) en Hambleton & Jones (1992) gaan ervan uit dat er bij een item sprake is van itembias, wanneer het item zowel bij allochtonen 1 versus autochtonen 1 als bij allochtonen 2 versus autochtonen 2 partijdig is.

Ten derde moeten er indices beschikbaar zijn die aangeven of er wel of niet sprake is van itembias en in welke mate dat het geval is. Verhelst (1992) geeft hiervoor een model.
Eerst wordt de kans op het juist beantwoorden van een item (item i) bepaald, gegeven een bepaalde totaalscore (score s) op de toets waar het item deel van uitmaakt. Deze kans, die een functie is van de itemparameter, kan geschreven worden als

$\text{Prob}(X_i=1 \mid \text{score}=s)$ of als $\pi_{i|s}$

Het aantal personen in niveaugroep s wordt geschreven als n_s.

Voor elke niveaugroep s wordt berekend welke proportie personen het juiste antwoord gegeven heeft op item i. Deze proportie wordt geschreven als

$p_{i|s}$. Als het IRT-model past dan is $\pi_{i|s} \approx p_{i|s}$.

Als het aantal personen in niveaugroep s klein is, dan worden twee of meer niveaugroepen samengevoegd tot de gecombineerde niveaugroep G. Aan de hand van de volgende (vereenvoudigde) formule kan voor elke subgroep de χ^2 bepaald worden.

$$\chi^2 = \sum_{q=1}^{r} \frac{\left[\sum_{Gq} n_s (p_{i|s} - \pi_{i|s})\right]^2}{\sum_{Gq} n_s \pi_{i|s} (1 - \pi_{i|s})}$$

De som van de χ^2 van de beide subgroepen geeft de χ^2-toetsingsgrootheid voor itembias aan. Het aantal vrijheidsgraden is gelijk aan het aantal niveaugroepen allochtone leerlingen plus het aantal niveaugroepen autochtone leerlingen minus twee (Verhelst, 1992).
Door de ICC's van allochtone en autochtone leerlingen ten opzichte van elkaar te onderzoeken, kan vastgesteld worden ten nadele van welke subgroep het item partijdig is.

Tot slot wordt er op gewezen dat het mogelijk is dat onder het gekozen IRT-model de items niet op een eendimensionele schaal blijken te passen. Mellenbergh (1989) en Glas (1991) geven aan dat dit laatste probleem op te lossen is door via een iteratieve procedure items te verwijderen die niet op de schaal passen. Het kan voorkomen dat een aantal niet-schaalbare items wel blijken te passen op een tweede schaal. Een verzameling items kan twee of meer afzonderlijke eendimensionale schalen bevatten, die dan afzonderlijk op itembias kunnen worden onderzocht. Het blijft evenwel mogelijk dat één of enkele items binnen een IRT-model niet op itembias onderzocht kunnen

worden, omdat ze in het geheel niet schaalbaar blijken te zijn (Glas, 1991). Glas & Verhelst (1993) en Shealy & Stout (1993) geven aan dat er ook multidimensionele IRT-modellen zijn waarmee bepaald kan worden in welke mate elk item uit een toets een beroep doet op twee of meer latente vaardigheden. Zoals in 1.2.2 reeds is opgemerkt, zijn deze relatief nieuwe modellen wiskundig ingewikkeld en de bruikbaarheid ervan voor onderzoek naar itembias is vooralsnog beperkt.

6.1.3 Opzet van de itembiasanalyses

Voor het onderzoek naar itembias is het tweede bestand gebruikt, dat bestaat uit gegevens van leerlingen waarvan zowel Eindtoets- als vragenlijstgegevens beschikbaar zijn (zie Hoofdstuk 4, tabel 4.3). In dit bestand varieert het aantal waarnemingen per etnische minderheidsgroep van 39 – 919. Eerst moet vastgesteld worden welke etnische groepen groot genoeg zijn voor onderzoek naar itembias. Intraprasert (1986) komt na onderzoek van vijf itembiasdetectie-procedures tot de conclusie dat een aantal van 400 – 500 waarnemingen per steekproef bij elke methode tot betrouwbare resultaten leidt. Bij Educational Testing Service (ETS) in de Verenigde Staten geldt als regel dat voor alle detectieprocedures bij voorkeur 500 waarnemingen per subgroep beschikbaar moeten zijn (Zieky, 1993). In verband hiermee is in het onderhavige onderzoek ervoor gekozen de Turkse (in 1987 = 797 en in 1989 = 919 leerlingen) en Marokkaanse leerlingen (in 1987 = 720 en in 1989 = 907 leerlingen) als onderzoeksgroep te laten fungeren. Als referentiegroep is een steekproef uit de autochtone leerlingen genomen (zie tabel 4.3). In verband met de controle op stabiliteit is het gewenst de analyses telkens uit te voeren op twee steekproeven uit elke subgroep (Mellenbergh, 1989; Bügel & Glas, 1991).

In 6.1 is reeds aangegeven dat een itembiasdetectieprocedure die gebaseerd is op de IRT en één die gebaseerd op de klassieke testtheorie zullen worden gehanteerd. In dit onderzoek is het computerprogramma 'One Parameter Logistic Model' (OPLM) gebruikt voor itembiasonderzoek onder het IRT-model (Verhelst, 1992). Als klassieke testtheorieprocedure is het Mantel-Haenszel-programma gebruikt (Verhelst, 1988). Zowel bij OPLM als bij het Mantel-Haenszel-programma is de 1% significantiegrens gehanteerd. Deze relatief ruime grens is gekozen in verband met de beperkte trefzekerheid bij het detecteren van partijdige items. Bovendien is het niet ongebruikelijk om de 1% significantiegrens in itembiasonderzoek te hanteren (Hambleton & Rogers, 1989; Bügel & Glas, 1991; Hambleton & Jones, 1992).

Omdat de Eindtoets Basisonderwijs samengesteld en geanalyseerd wordt volgens de klassieke testtheorie wordt eerst *de Mantel-Haenszel-procedure* gehanteerd (zie 6.1.1). Van de Eindtoets Basisonderwijs 1987 en 1989 wordt voor de items van de onderdelen Taal, Rekenen en Informatieverwerking (per jaar: 60 + 60 + 60 = 180 items) telkens vastgesteld welke items partijdig zijn ($p<.01$). Bij de eerste analyse zijn alle 60 items van een onderdeel opgenomen in de totaalscore, daarna is de totaalscore 'gezuiverd' van partijdige items.
Bij de tweede analyse zijn alleen de items uit de totaalscore verwijderd die bij de eerste analyse in grote mate partijdig waren. Bij de volgende analyse zijn alle partijdige items uit de totaalscore verwijderd. Dan blijken er soms nieuwe

partijdige items bij te komen, maar het is ook mogelijk dat items niet meer partijdig zijn die bij een vorige analyse wel partijdig waren. Het iteratieve proces zal zo veel mogelijk doorgaan totdat de totaalscore gebaseerd kan worden op een verzameling onpartijdige items. Het is echter niet mogelijk om, indien nodig, onbeperkt items uit de totaalscore te verwijderen, omdat de overgebleven items het domein voldoende moeten representeren. In verband met de inhoudsvaliditeit is er naar gestreefd om per toetsonderdeel de totaalscore op minstens tweederde van het totaal aantal items te blijven baseren. Als er éénderde deel van de items verwijderd wordt, dan is dat geen a-selecte steekproef uit het totaal aantal items. Er zal nagegaan worden in welke mate de overgebleven items het domein voldoende representeren.

Bij de Mantel-Haenszel-analyses zal vastgesteld moeten worden in hoeverre er sprake is van niet-uniforme itembias (Uiterwijk, 1990a). Bij niet-uniforme itembias zijn items partijdig bij laagpresterende niveaugroepen en niet bij hoogpresterende of omgekeerd. Met het Mantel-Haenszel-programma (Verhelst, 1988) is het alleen mogelijk om uniforme itembias te traceren en geen niet-uniforme bias. Om aanwijzingen te verkrijgen voor niet-uniforme itembias zijn afzonderlijke Mantel-Haenszel-analyses uitgevoerd voor laag- en hoogscorende leerlingen. Om ook een externe maat voor het onderscheid hoog-, respectievelijk laagpresterende kinderen te nemen, zijn tevens Mantel-Haenszel-analyses uitgevoerd met betrekking tot leerlingen met een hoog, respectievelijk laag advies basisschool. Zowel bij 'score' als bij 'advies' is de grens tussen laag- en hoogpresterende leerlingen zo gekozen, dat het aantal allochtone leerlingen in beide groepen ongeveer even groot is.
Omdat de analyses met hoog-, respectievelijk laagpresterende leerlingen eveneens de mogelijkheid bieden de stabiliteit van de Mantel-Haenszel-analyses te onderzoeken zijn geen afzonderlijke analyses op a-selecte steekproeven uitgevoerd. De stabiliteit van de Mantel-Haenszel-analyses kan blijken door de z-waarden van de analyses met hoog- en laagpresterende leerlingen te vergelijken.
Bij het Mantel-Haenszel-computerprogramma (Verhelst, 1988) moet per analyse opgegeven worden wat het minimum aantal leerlingen per niveaugroep is. Om de fluctuaties bij verschillende aantallen na te gaan, zijn bij hetzelfde leerlingenbestand en toetsonderdeel analyses verricht met 25, 50, 100, 200 en 400 leerlingen. De correlaties tussen de z-waarden bleken niet lager dan .99 te zijn. Wel bleken enkele items bij de ene analyse wel en bij de volgende analyse niet partijdig te zijn ($p<.01$). Om de stabiliteit van de analyses op dit punt te vergroten is er naar gestreefd om het aantal leerlingen per niveaugroep zo te nemen dat het aantal leerlingen in de onderzoeksgroep, respectievelijk de referentiegroep, niet lager is dan vijf.

Voor de *analyses onder het IRT-model* is het niet mogelijk om te starten met alle 60 items van een onderdeel, omdat niet te verwachten is dat deze items een eendimensionele schaal vormen (Uiterwijk, 1990a). Eerst zal bepaald worden welke schalen in elk toetsonderdeel onderscheiden moeten worden.

Omdat de Eindtoets Basisonderwijs volgens de klassieke testtheorie wordt samengesteld en geanalyseerd ligt het voor de hand dat de items van de toets-onderdelen Taal, Rekenen en Informatieverwerking niet voldoen aan het

Raschmodel. Het is te verwachten dat met dit model een deel van de items niet schaalbaar blijken te zijn en in feite niet onderzocht kunnen worden op itembias (vgl. Glas, 1991). In verband hiermee is bij het computerprogramma OPLM naast de moeilijkheidsparameter ook als hypothese een discriminatie-index ingevoerd. Vervolgens wordt statistisch getoetst of de discriminatie-index aanvaard kan worden (Verhelst, 1992). Dit heeft tot gevolg dat het item in zijn bijdrage aan de totaalscore niet het gewicht één krijgt, zoals onder het Raschmodel, maar dat het item wordt gewogen met zijn discriminatie-index. Hierbij is met betrekking tot de onderdelen Taal, Rekenen en Informatie-verwerking van de Eindtoets Basisonderwijs 1987 en 1989 telkens de volgende procedure gehanteerd (zie Mellenbergh, 1989; Glas, 1991).

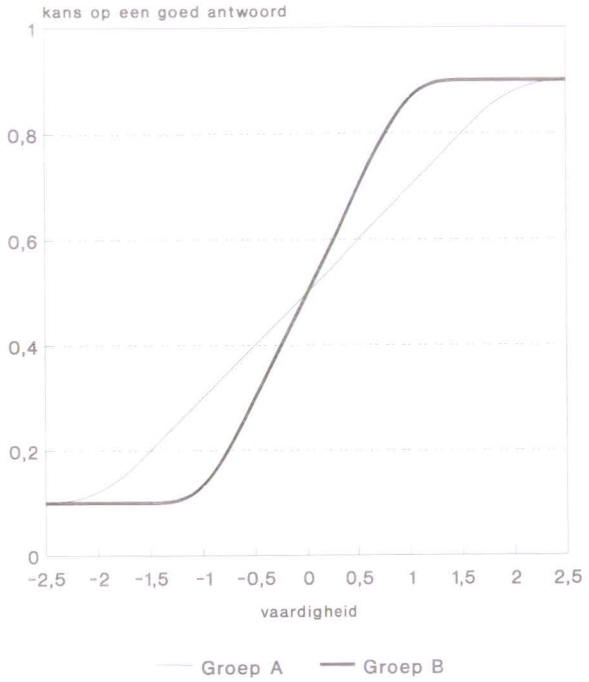
- De autochtone, Turkse en Marokkaanse leerlingen zijn a-select in twee subgroepen verdeeld.
- Er is onderzocht welke eendimensionele schalen elk toetsonderdeel bevat voor autochtone leerlingen. Hiervoor is eerst per toetsonderdeel vastgesteld welke factorstructuur er in de tetrachorische correlaties tussen de items te onderkennen is. Daarna is onderzocht of de items die op een bepaalde factor hoog laden daadwerkelijk op een eendimensionele schaal passen. Een schaal is als zodanig aanvaard wanneer de items op een schaal passen en de discriminatie-index van elk item in beide steekproeven (autochtonen 1 en autochtonen 2) gelijk is.
- Een schaal is als een 'baseline' voor onderzoek naar itembias geaccepteerd wanneer er geen enkel item in de vergelijking autochtonen 1 versus autochtonen 2 partijdig is ($p<.01$)
- Voor elke steekproef autochtone versus Turkse, respectievelijk Marokkaanse leerlingen is vastgesteld welke items partijdig zijn ($p<.01$).

Hierbij zijn per toetsonderdeel de volgende analyses verricht:

- Autochtonen 1 versus Turkse leerlingen 1
- Autochtonen 2 versus Turkse leerlingen 2
- Autochtonen 1 versus Marokkaanse leerlingen 1
- Autochtonen 2 versus Marokkaanse leerlingen 2

Ook bij op IRT gebaseerde procedures zal onderzocht worden of er sprake is van niet-uniforme itembias. Door inspectie van de ICC's kan vastgesteld worden of er sprake is van niet-uniforme itembias. In dat geval zal zoals in figuur 6.3 is aangegeven, de ICC van de ene groep de ICC van de andere groep kruisen.

Figuur 6.3 De 'item characteristic curves' van twee groepen bij niet-uniforme itembias



Nadat de items van de Eindtoets Basisonderwijs 1987 en 1989 met OPLM zijn onderzocht, zijn de items per schaal ook nog eens met Mantel-Haenszel-procedure geanalyseerd. Hierdoor is het mogelijk de resultaten van de OPLM-analyses met die van de Mantel-Haenszel-analyses te vergelijken. De totaalscore bevat bij deze analyses de items die bij de twee steekproeven autochtone leerlingen een schaal vormen. De totaalscore is bij deze Mantel-Haenszel-analyses niet 'gezuiverd', omdat deze items voor de autochtone populatie onpartijdig zijn en dit dus in feite ook bij de allochtone leerlingen zouden moeten zijn. In het onderhavige onderzoek is de Mantel-Haenszel-procedure dus twee keer gebruikt:

- Analyses waarbij in eerste instantie de 60 items van een toetsonderdeel in de totaalscore zijn opgenomen. De totaalscore is daarna zoveel mogelijk 'gezuiverd' van partijdige items.
- Analyses waarbij alle items van de IRT-schaal de gewogen totaalscore vormen.

6.2 Resultaten van de analyses naar itembias

Zoals eerder vermeld, zijn de items van de Eindtoets Basisonderwijs 1987 en 1989 met de Mantel-Haenszel-procedure en met een op het IRT-model gebaseerde procedure op itembias voor allochtone leerlingen onderzocht. De uitkomsten van de Mantel-Haenszel-analyses worden besproken in 6.2.1. De IRT-analyses zijn uitgevoerd op de eendimensionele schalen die in elk toetsonderdeel gevonden zijn. Op dezelfde schalen zijn ook Mantel-Haenszel-

analyses uitgevoerd. De resultaten van deze analyses worden besproken in 6.2.2.

6.2.1 De resultaten van de Mantel-Haenszel-analyses

Bij de Mantel-Haenszel-procedure worden de leerlingen met de totaalscore in niveaugroepen ingedeeld. Met een iteratieve procedure is de totaalscore zoveel mogelijk gezuiverd van partijdige items. De items die bij de eerste analyse partijdig blijken te zijn, worden bij de tweede analyse niet in de totaalscore opgenomen. Zoals in 6.1.3 is gezegd, is in verband met de inhoudsvaliditeit van de totaalscore niet meer dan éénderde van het aantal items verwijderd. Dit heeft tot gevolg dat de totaalscore bij een deel van de analyses op partijdige items is gebaseerd.
Het kwam ook voor dat een item bij de eerste analyse partijdig is en dat bij de tweede analyse niet meer is. Wanneer een dergelijk item bij de derde analyse weer in de totaalscore wordt opgenomen, is het meestal wel weer partijdig. Dat hetzelfde item de ene keer wel en de andere keer niet partijdig is, wordt veroorzaakt door de totaalscore, die niet steeds op dezelfde verzameling items is gebaseerd. De op verschillende manieren samengestelde totaalscores representeren niet telkens dezelfde vaardigheid. Dit geeft aanleiding te veronderstellen dat de items van de Eindtoetsonderdelen geen eendimensionele vaardigheid vertegenwoordigen.
De items die partijdig zijn in het voordeel van allochtone leerlingen hebben een z–waarde <-2.58, de items die in het nadeel van allochtone leerlingen partijdig zijn hebben een z-waarde >2.58. In de tabellen 6.1 – 6.3 staan de afgeleide gegevens van de uitgevoerde Mantel-Haenszel-analyse. De gemiddelde z-waarden en de standaarddeviaties in deze tabellen zijn gebaseerd op de z-waarden van de 60 items van het betreffende toetsonderdeel.

Tabel 6.1 Aantal partijdige taalitems met de Mantel-Haenszel-procedure waarbij het 'gezuiverde' toetsonderdeel als totaalscore geldt

Groep	Aantal part. items	Voor allochtonen in		Gem. z-waarde	sd
		nadeel	voordeel		
1987 (k = 60)					
Alle Turken	34	22	12	2.59	5.58
Hoge score	31	19	12	1.43	4.67
Lage score	8	8	0	0.53	2.45
Hoog advies	31	22	9	1.77	4.29
Laag advies	16	15	1	1.47	3.59
Alle Marokkanen	31	25	6	2.52	4.57
Hoge score	24	15	9	0.97	3.55
Lage score	6	5	1	0.27	2.04
Hoog advies	20	18	2	1.48	3.42
Laag advies	19	17	2	1.31	2.86
1989 (k = 60)					
Alle Turken	32	21	11	1.72	4.81
Hoge score	27	17	10	0.85	3.99
Lage score	19	17	2	1.16	2.60
Hoog advies	25	18	7	1.45	4.11
Laag advies	18	17	1	1.47	3.04
Alle Marokkanen	26	18	8	1.22	4.36
Hoge score	16	11	5	0.93	3.62
Lage score	14	8	6	0.36	2.64
Hoog advies	20	14	6	1.00	3.69
Laag advies	17	15	2	0.95	2.75

Er zijn in 1987 zeven taalitems (nr. 11, 12, 14, 26, 29, 31 en 34) die voor alle onderscheiden Turkse subgroepen partijdig zijn; de taalitems 11, 14, 26 en 31 zijn ook voor de Marokkaanse leerlingen partijdig. In 1989 zijn er 13 items (nr. 4, 7, 12 t/m 14, 16, 17, 24, 26, 31, 34, 39 en 55) partijdig voor alle Turkse subgroepen; de items 16, 26, 31, 34, 39 zijn ook steeds partijdig voor de Marokkaanse leerlingen. Taalitem 52 is alleen bij de analyses van Marokkaanse leerlingen partijdig. In hoofdstuk zeven wordt ingegaan op de vraag wat bij deze items de oorzaak van itembias zou kunnen zijn.
Uit nadere analyse is gebleken dat bij de Turkse en Marokkaanse leerlingen de 'gezuiverde' taaltotaalscore door de iteratieve procedure meer van taalgebruik-items is gezuiverd dan van spellingitems. De taalvaardigheidscore die gebruikt is om de leerlingen in niveaugroepen in te delen bestaat dus in feite voor een groter deel uit spellingitems dan de 'ongezuiverde' taaltotaalscore. Dit geeft aanleiding te veronderstellen dat de 60 taalitems voor autochtone leerlingen en/of voor Turkse, respectievelijk Marokkaanse leerlingen een multi-dimensionele schaal vormen. Bij de analyses onder het IRT-model (zie 6.2.2) zal

nagegaan worden in hoeverre het onderscheid taalgebruik- en spellingitems in de vorm van afzonderlijke eendimensionele schalen terug te vinden is.

Tabel 6.2 Aantal partijdige rekenitems met de Mantel-Haenszel-procedure waarbij het 'gezuiverde' toetsonderdeel als totaalscore geldt

Groep	Aantal part. items	Voor allochtonen in		Gem. z-waarde	sd
		nadeel	voordeel		
1987 (k = 60)					
Alle Turken	20	14	6	1.12	3.01
Hoge score	15	10	5	0.68	2.73
Lage score	5	5	0	0.23	1.68
Hoog advies	11	11	0	1.01	2.56
Laag advies	4	4	0	0.22	1.78
Alle Marokkanen	11	8	3	0.50	2.35
Hoge score	9	8	1	0.43	2.12
Lage score	6	5	1	0.22	1.51
Hoog advies	10	9	1	0.76	2.15
Laag advies	4	4	0	0.16	1.46
1989 (k = 60)					
Alle Turken	15	12	3	0.68	2.56
Hoge score	14	12	2	0.77	2.42
Lage score	6	4	2	0.16	1.66
Hoog advies	11	11	0	0.73	2.16
Laag advies	5	3	2	-0.01	1.75
Alle Marokkanen	12	8	4	0.39	2.38
Hoge score	8	2	6	-0.24	1.94
Lage score	8	8	0	0.47	1.62
Hoog advies	8	5	3	0.10	1.93
Laag advies	4	3	1	0.15	1.65

Er zijn in 1987 twee rekenitems (nr. 27 en 29) die voor alle onderscheiden Turkse subgroepen partijdig zijn; het rekenitem 27 is ook voor de Marokkaanse leerlingen partijdig. Item 45 is alleen voor de Marokkaanse subgroepen partijdig. In 1989 zijn er ook twee items (nr. 27 en 28) partijdig voor alle Turkse subgroepen; item 28 is ook steeds partijdig voor de Marokkaanse leerlingen. In hoofdstuk zeven zal ingegaan worden op de vraag wat bij deze items de itembias zou kunnen veroorzaken.

Tabel 6.3 Aantal partijdige informatieverwerkingitems met de Mantel-Haenszel-procedure waarbij het 'gezuiverde' toetsonderdeel als totaalscore geldt

Groep	Aantal part. items	Voor allochtonen in		Gem. z-waarde	sd
		nadeel	voordeel		
1987 (k = 60)					
Alle Turken	21	17	4	1.28	3.77
Hoge score	13	10	3	0.74	3.11
Lage score	13	8	5	0.13	2.10
Hoog advies	16	14	2	1.19	2.90
Laag advies	13	9	4	0.40	2.61
Alle Marokkanen	23	18	5	1.39	3.38
Hoge score	19	12	7	0.36	2.64
Lage score	7	6	1	0.21	1.92
Hoog advies	8	6	2	0.30	2.24
Laag advies	14	11	3	0.55	2.41
1989 (k = 60)					
Alle Turken	18	14	4	1.09	3.07
Hoge score	15	13	2	0.84	2.54
Lage score	7	7	0	0.15	1.88
Hoog advies	17	16	1	1.16	2.62
Laag advies	9	8	1	0.44	2.05
Alle Marokkanen	21	20	1	1.85	3.17
Hoge score	15	14	1	0.96	2.42
Lage score	6	3	3	0.03	1.80
Hoog advies	17	17	0	1.46	2.83
Laag advies	6	6	0	0.30	1.80

Er zijn in 1987 vier informatieverwerkingitems (nr. 13, 16, 26 en 31) die voor alle onderscheiden Turkse subgroepen partijdig zijn; de informatieverwerking-items 16, 26 en 31 zijn ook voor de Marokkaanse leerlingen partijdig. Item 29 is alleen voor de Marokkaanse subgroepen partijdig. In 1989 zijn er vijf items (nr. 9, 15, 19, 20, 47) partijdig voor alle Turkse subgroepen; de items 9 en 19 zijn ook steeds partijdig voor de Marokkaanse leerlingen. In hoofdstuk zeven wordt ingegaan op de vraag wat bij deze items de oorzaak van itembias zou kunnen zijn.

Wanneer we de tabellen 6.1 – 6.3 als een geheel beschouwen dan blijkt dat we moeilijk kunnen aangeven in welke mate de Eindtoets Basisonderwijs 1987 en 1989 partijdige items bevatten.
Wanneer we naar de gemiddelde z-waarden kijken dan blijkt dat, afgezien van het taalonderdeel uit de Eindtoets Basisonderwijs 1987, de gemiddelde z-waarden lager zijn dan 2.58. Bij het rekenonderdeel van de Eindtoets Basisonderwijs 1989 komen zelfs twee negatieve gemiddelde z-waarden voor.

Wanneer we naar de kolommen met het aantal partijdige items kijken, dan zien we een wisselend beeld. Het onderdeel Taal bestaat voor alle Turkse en Marokkaanse leerlingen voor ongeveer de helft uit partijdige items, het onderdeel Rekenen voor éénderde tot éénvijfde deel en Informatieverwerking voor omstreeks éénderde deel. Bij vrijwel alle analyses is het aantal partijdige items in het nadeel van allochtone leerlingen groter dan het aantal items in het voordeel. Een uitzondering vormen de rekenitems 1989 voor hoogscorende Marokkaanse leerlingen.

Het aantal partijdige items is altijd lager bij de analyses met leerlingen met een lage toetsscore of een laag advies. Uit nadere analyse blijkt dat het in 1987 bij 8.3% van de items voorkomt dat een item bij leerlingen van een laag (2.2%), respectievelijk hoog (6.1%) prestatieniveau partijdig is en niet bij totale groep Turkse of Marokkaanse leerlingen. In 1989 is dit bij 7.8% van de items het geval; 2.0% voor leerlingen van een laag prestatieniveau en 5.8% bij een hoog prestatieniveau. Hieruit blijkt dat de gehanteerde Mantel-Haenszel-procedure (Verhelst, 1988) zoals verwacht alleen uniforme itembias kan opsporen en geen niet-uniforme. Bij de uitgevoerde analyses op Eindtoets Basisonderwijs 1987 en 1989 blijkt dit dus in omstreeks 8% van de items het geval te zijn.
Uit de tabellen 6.1 – 6.3 blijkt eveneens dat het aantal items in het nadeel van Turkse en Marokkaanse leerlingen hoger is en het aantal items in het voordeel lager, wanneer het (hoog dan wel laag) advies van de basisschool gebruikt wordt om de leerlingen in niveaugroepen in te delen. Een uitzondering op deze regel vormt Rekenen 1989.

Over het geheel genomen blijkt uit de tabellen 6.1 – 6.3 dat Taal het grootste aantal partijdige items bevat gevolgd door respectievelijk Informatieverwerking en Rekenen. De drie toetsonderdelen bevatten meestal minder partijdige items voor Marokkaanse dan voor Turkse leerlingen. Bij Taal zijn er meer items in het voordeel van Turkse dan van Marokkaanse leerlingen; dit geldt ook voor de rekenitems uit 1987 en de informatieverwerkingitems uit 1989. Over het algemeen bevat de Eindtoets Basisonderwijs 1989 minder partijdige items dan de toets uit 1987.
De resultaten van de Mantel-Haenszel-procedure laten slechts voorlopige conclusies toe. Tot nu toe is nog niet vastgesteld of de items van een toetsonderdeel een eendimensionele schaal vormen. Het is niet uitgesloten dat een aantal items partijdig is, omdat ze zowel voor autochtone als voor allochtone leerlingen multidimensioneel zijn. Er moet nog vastgesteld worden in welke mate de IRT-analyses het beeld van de Mantel-Haenszel-analyses bevestigen.

6.2.2 De resultaten van de IRT-analyses

Om de stabiliteit van de analyses naar itembias te controleren zijn eerst de autochtone, de Turkse en de Marokkaanse leerlingen a-select in twee gelijke steekproeven verdeeld (zie 6.1.3). Voordat op basis van de itemresponsen van de twee steekproeven de opgaven van de Eindtoets Basisonderwijs 1987 en 1989 met een IRT-model konden worden onderzocht, moest eerst vastgesteld worden welke eendimensionele schalen elk toetsonderdeel bevat. De procedure die hiervoor is gevolgd, staat vermeld in 6.1.3. Elk toetsonderdeel bleek afgezien

van Rekenen in 1989 uit twee eendimensionele schalen te bestaan. Het rekenonderdeel uit 1989 bevat drie schalen. De items uit schaal Taal B hebben in beide jaren voornamelijk betrekking op spelling, de items uit Taal A op taalgebruik (zie tabel 3.1). De informatieverwerkingitems uit schaal A zijn vrijwel allemaal afkomstig uit de opgavenrubriek Lezen van teksten, terwijl de items uit schaal B bijna helemaal uit de opgavenrubrieken Hanteren van informatiebronnen, Kaartlezen en Lezen van tabellen en grafieken afkomstig zijn. De rekenschalen zijn minder eenvoudig te kenmerken.

Nadat is vastgesteld welke schalen de Eindtoets Basisonderwijs 1987 en 1989 bevatten, zijn met de IRT- en Mantel-Haenszel-procedure de items per schaal op itembias onderzocht. Door de beide itembiasdetectieprocedures te hanteren worden de verschillende effecten van beide technieken zichtbaar.
Voor de duidelijkheid wordt nog opgemerkt dat het verschil tussen de Mantel-Haenszel-analyses van 6.2.1 en 6.2.2 gelegen is in het feit dat in 6.2.1 de totaalscore (na 'zuivering') gebaseerd is op de items van een toetsonderdeel, in 6.2.2 is de totaalscore gebaseerd op de items die op een eendimensionele schaal passen. De wijze waarop de totaalscore bij de Mantel-Haenszel- en bij de IRT-procedure berekend wordt, is niet gelijk. Bij de Mantel-Haenszel-techniek wordt de totaalscore bepaald door de som van het aantal goed beantwoorde items. Bij het gehanteerde IRT-model levert elk goed beantwoord item een gewogen score op en is de totaalscore de som van de gewogen scores. Door het computerprogramma OPLM wordt ook aangegeven wat de correlatie tussen de gewogen en ongewogen scores is. Aangezien bij de uitgevoerde analyses deze correlatie altijd .98 of hoger is, zijn de verschillen tussen gewogen en ongewogen scores klein te noemen.

In Tabel 6.4 staan per schaal de itemnummers van de Eindtoets Basisonderwijs 1987 en 1989 vermeld die partijdig zijn zowel in steekproef 1 als in steekproef 2. Een item kan partijdig zijn met de IRT-analyse, met de Mantel-Haenszel-analyse (MH) of met beide.
In hoofdstuk zeven wordt besproken waarom de in tabel 6.4 genoemde items partijdig kunnen zijn.

*Tabel 6.4 Nummers van partijdige items met de Mantel-Haenszel- en de IRT-procedure **

Groep/schaal	MH	IRT	MH & IRT
1987			
Turken			
Taal A (k=36)	26	**51**, **54**	11, 12, 14
Taal B (k=24)	**46**	–	28, 31, **47**
Rekenen A (k=23)	29, 45	–	–
Rekenen B (k=37)	41, 57	–	–
Info A (k=34)	4, 13, **49**, 52	**55**	2
Info B (k=26)	16, **23**, **33**	–	–
Marokkanen			
Taal A (k=36)	14	**15**	11
Taal B (k=24)	28, **46**, **55**, **60**	–	31
Rekenen A (k=23)	29, 45	–	–
Rekenen B (k=37)	–	–	–
Info A (k=34)	–	**55**	2
Info B (k=26)	–	–	–
1989			
Turken			
Taal A (k=40)	17, **25**, 31, 34, **35**, 39, **41**	–	**16**, 26
Taal B (k=20)	50, 55, **57**	–	–
Rekenen A (k=14)	–	–	–
Rekenen B (k=13)	–	–	–
Rekenen C (k=33)	–	–	–
Info A (k=36)	20, **44**	–	47
Info B (k=24)	19	–	**1**
Marokkanen			
Taal A (k=40)	**21**, 28	31	**16**, 26, 34
Taal B (k=20)	55	–	–
Rekenen A (k=14)	–	32	–
Rekenen B (k=13)	–	–	–
Rekenen C (k=33)	–	–	–
Info A (k=36)	20	47	–
Info B (k=24)	–	–	**1**, 19

* *De vetgedrukte items zijn partijdig in het voordeel van allochtone leerlingen*

Uit Tabel 6.4 blijkt dat met zowel de Mantel-Haenszel- als de IRT-procedure er 19 items van de Eindtoets Basisonderwijs 1987 en 1989 partijdig zijn bij Turkse en/of Marokkaanse leerlingen. Van die 19 items zijn er zes zowel bij Turkse als Marokkaanse leerlingen partijdig. De Eindtoets Basisonderwijs 1987 en 1989 bevatten dus in totaal 13 items die met zowel de Mantel-Haenszel- als de IRT-procedure partijdig zijn bij één of beide etnische minderheidsgroepen. Van deze 13 items zijn er drie partijdig in het voordeel van allochtone leerlingen, tien in

het nadeel. Verder blijkt dat acht items van de Eindtoets Basisonderwijs 1987 en 1989 partijdig zijn bij de IRT-procedure; één item daarvan is zowel bij Turkse als bij de Marokkaanse leerlingen partijdig. Van deze zeven items zijn er vier in het voordeel van allochtone leerlingen en drie in het nadeel. Uit Tabel 6.4 blijkt dat met de Mantel-Haenszel-procedure er 37 items partijdig zijn bij Turkse en/of Marokkaanse leerlingen. Van die 37 items zijn er vijf zowel bij Turkse als Marokkaanse leerlingen partijdig. De Eindtoets Basisonderwijs 1987 en 1989 bevatten dus in totaal 32 items, die met de Mantel-Haenszel-procedure partijdig zijn bij één of beide etnische minderheidsgroepen. Van deze 32 items zijn er 12 in het voordeel en 20 in het nadeel van allochtone leerlingen. Verder blijkt uit Tabel 6.4 dat partijdige items nooit in het voordeel van Turkse leerlingen zijn en in het nadeel van Marokkaanse leerlingen of omgekeerd.

Uit Tabel 6.4 ontstaat misschien ten onrechte de indruk dat de resultaten van de beide itembiasdetectieprocedures aanzienlijk verschillen. Er zijn in totaal 360 items voor Turkse leerlingen en dezelfde 360 items zijn ook voor Marokkaanse leerlingen geanalyseerd: 720 afzonderlijke analyses. Zoals gezegd zijn zowel de Turkse als de Marokkaanse leerlingen in twee a-selecte steekproeven verdeeld. Een item wordt in tabel 6.4 als partijdig aangemerkt als het item in beide steekproeven partijdig is ($p<.01$). Uit nadere analyses is gebleken dat de Mantel-Haenszel-techniek in 14% van de analyses wel in de ene steekproef maar niet in de andere steekproef partijdige items detecteert; er is in 86% van de analyses dus sprake van overeenstemming. Bij de IRT-procedure is er in 89% van de analyses overeenstemming tussen beide steekproeven. Deze percentages hebben dus betrekking op de stabiliteit van beide soorten analyses.

Wanneer we naar de overeenstemming tussen beide procedures binnen elke steekproef afzonderlijk kijken dan blijkt dat de beide procedures in 87% van de analyses overeenstemmen; dit percentage geldt zowel voor 1987 als voor 1989. Wanneer de beide procedures afwijken, dan wordt dat in 1987 voor 73% van de gevallen veroorzaakt door de Mantel-Haenszel-procedure, in 1989 ligt dit percentage op 70%. Opmerkelijk is dat Hambleton & Jones (1992) bij een vergelijking tussen de resultaten van de Mantel-Haenszel- en een IRT-procedure een overeenstemming van 88% constateerden. Bügel & Glas (1991) komen uit op 91% overeenstemming. Bij het IRT-model dat Hambleton & Jones en Bügel & Glas hanteren (Rasch-model) is de totaalscore niet de som van de gewogen scores, maar de som van de ongewogen scores. In het onderhavige onderzoek is het verschil tussen beide soorten totaalscores klein te noemen, omdat de correlatie tussen gewogen en ongewogen scores nooit lager is dan .98.

Er zijn dus 360 Eindtoetsitems door twee itembiasdetectieprocedures voor zowel Turkse als Marokkaanse leerlingen geanalyseerd. Van die 360 items worden er 52 items (=14%) één of twee keer in tabel 6.4 genoemd. Van de onderzochte 360 items is dus 86% niet in twee steekproeven partijdig.

In 6.2.1 zijn de resultaten opgenomen van de Mantel-Haenszel-analyses met het 'gezuiverde' toetsonderdeel als totaalscore. In 6.2.2 staan de resultaten van zowel de Mantel-Haenszel- als de IRT-procedure met de items van de eendimensionele schaal als totaalscore. Bij vergelijking van de resultaten van

6.2.1 en 6.2.2 wordt duidelijk dat het moeilijk is om aan te geven of een item nu wel of niet partijdig is.
Bij de Mantel-Haenszel-analyses is het van belang of de totaalscore gebaseerd is op het 'gezuiverde' toetsonderdeel of op een eendimensionele schaal. Bij vergelijking van tabel 6.4 met de tabellen 6.1 – 6.3 blijkt dat de Eindtoets Basisonderwijs de ene keer meer partijdige items bevat dan de andere keer.

Tabel 6.5 Aantal partijdige items met de Mantel-Haenszel-procedure waarbij het 'gezuiverde' toetsonderdeel of de eendimensionele schaal als totaalscore geldt

	totaalscore: 'gezuiverde' toetsonderdeel	totaalscore: eendimensionele schaal
1987 (k=180)		
Turken	75 (=42%)	20 (=11%)
Marokkanen	65 (=36%)	10 (= 6%)
1989 (k=180)		
Turken	65 (=36%)	17 (= 9%)
Marokkanen	59 (=33%)	9 (= 5%)
Totaal (k=720)	264 (=37%)	56 (= 8%)

Uit tabel 6.5 blijkt dat de Mantel-Haenszel-techniek met het 'gezuiverde' toetsonderdeel als totaalscore bij dezelfde toetsitems aanzienlijk meer partijdige items oplevert dan met de eendimensionele schaal als totaalscore. Deze bevinding stemt overeen met die van Clauser e.a. (1991), die constateerden dat de Mantel-Haenszel-procedure 32% minder partijdige items opspoort wanneer de totaalscore niet gebaseerd is op alle items van een bepaald toetsonderdeel maar op de items van inhoudelijk meer samenhangende subtests.
Er bestaat aanzienlijke overlap tussen de beide Mantel-Haenszel-procedures, omdat de items die met de eendimensionele schaal als totaalscore partijdig zijn, op twee na (1989: Taal nr. 55 en Informatieverwerking nr. 1) ook partijdig zijn met het 'gezuiverde' toetsonderdeel als totaalscore. Zoals gezegd levert deze laatste procedure echter meer partijdige items op.
Van de 20 items die met de IRT-procedure partijdig zijn, zijn er 15 ook partijdig met de Mantel-Haenszel-procedure met het 'gezuiverde' toetsonderdeel als totaalscore.
Over het algemeen kan gesteld worden dat de Mantel-Haenszel-techniek meer partijdige items opspoort dan de IRT-procedure. Gebleken is dat de beide technieken dit ongeveer even stabiel doen: de Mantel-Haenszel-techniek stemt bij twee steekproeven in 86% van de analyses overeen, de IRT-techniek doet dit in 89% van de gevallen. Op basis van deze gegevens kan niet bepaald worden welke itembiasdetectieprocedure de juiste is.

6.3 Samenvatting en conclusie

Een item is partijdig wanneer leerlingen uit verschillende subgroepen maar met hetzelfde vaardigheidsniveau een ongelijke kans hebben om het betreffende item goed te beantwoorden. Voor het opsporen van partijdige items in de Eindtoets Basisonderwijs 1987 en 1989 voor allochtone leerlingen zijn twee itembiasdetectieprocedures gebruikt (zie 6.1). Het computerprogramma One Parameter Logistic Model (OPLM) (Verhelst, 1992) is gebruikt als procedure die gebaseerd is op de itemresponsetheorie (IRT); het Mantel-Haenszel-programma (Verhelst, 1988) is gehanteerd als procedure gebaseerd op de klassieke testtheorie. Het Mantel-Haenszel-programma gaat van de assumptie uit dat het totaal aantal goed gemaakte opgaven een juiste schatting is van de te meten vaardigheid. Onder het IRT-model wordt getoetst of deze aanname juist is door te onderzoeken of de items een eendimensionele schaal vormen.
In verband met het aantal waarnemingen per etnische groep zijn de items van de drie toetsonderdelen Taal, Rekenen en Informatieverwerking uit de Eindtoets Basisonderwijs 1987 en 1989 alleen voor Turkse en Marokkaanse in vergelijking met autochtone leerlingen op itembias onderzocht. Omdat de Eindtoets Basisonderwijs volgens de klassieke testtheorie samengesteld en geanalyseerd wordt, is gestart met de Mantel-Haenszel-procedure. De totaalscore van het toetsonderdeel die bij de Mantel-Haenszel-procedure wordt gehanteerd om de leerlingen in niveaugroepen in te delen, wordt via een iteratieve procedure zoveel mogelijk van partijdige items 'gezuiverd'. De items die bij de eerste analyse van elk toetsonderdeel partijdig blijken te zijn, worden bij een volgende analyse niet meer opgenomen in de totaalscore. Omdat het Mantel-Haenszel-programma alleen uniforme itembias kan opsporen (items met niet-uniforme itembias zijn niet over het hele totaalscorebereik partijdig) zijn ook afzonderlijke analyses uitgevoerd voor laag- en hoogpresterende leerlingen. Deze laatste analyses zijn ook verricht om de stabiliteit van de analyses te onderzoeken.
Vervolgens zijn de items uit de Eindtoets Basisonderwijs 1987 en 1989 onderzocht met een procedure gebaseerd op het IRT-model. Om de stabiliteit te onderzoeken zijn hierbij de analyses verricht op twee steekproeven uit autochtone, respectievelijk Turkse en Marokkaanse leerlingen. Eerst is met het eenparameter-model (moeilijkheidsgraadparameter en discriminatie-index) vastgesteld welke items een eendimensionele schaal vormen en als basis kunnen dienen voor onderzoek naar itembias. De items van elke eendimensionele schaal zijn met OPLM op itembias onderzocht en ook nog met de Mantel-Haenszel-procedure. Hierdoor is het mogelijk de resultaten van de Mantel-Haenszel- met die van de IRT-procedure te vergelijken.

De resultaten van de analyses naar itembias (zie 6.2) maken duidelijk dat het moeilijk is om aan te geven hoeveel items van de Eindtoets Basisonderwijs 1987 en 1989 partijdig zijn. De verschillende analyses laten een wisselend beeld zien. Wanneer we de Mantel-Haenszel-analyses met de 'gezuiverde' toetsonderdeelscore als maatstaf nemen dan is voor Turkse en voor Marokkaanse leerlingen zowel in 1987 als in 1989 ongeveer de helft van de 60 taalitems partijdig, de 60 rekenitems zijn voor éénderde tot éénvijfde deel partijdig en de 60 informatieverwerkingitems voor éénderde deel. De drie toetsonderdelen bevatten meestal minder partijdige items voor Marokkaanse dan voor Turkse

leerlingen. Het aantal partijdige items is over het geheel genomen nog 8% hoger in verband met het aantal items waarbij sprake is van niet-uniforme itembias. De resultaten van de Mantel-Haenszel-procedure met de 'gezuiverde' toetsonderdeelscore laten slechts voorlopige conclusies toe. Tot nu toe is nog niet vastgesteld of de items van een toetsonderdeel een eendimensionele schaal vormen, waardoor het niet uitgesloten is dat een aantal items partijdig is, omdat ze zowel voor autochtone als voor allochtone leerlingen multidimensioneel zijn. Bij de analyses met de IRT-procedure blijkt dat het aantal partijdige items voor Turkse en/of Marokkaanse leerlingen beduidend geringer is: 20 van de in totaal 360 geanalyseerde items (=6%). Opgemerkt moet worden dat bij de IRT-analyses de totaalscore gebaseerd is op de items van de eendimensionele schaal. Als we de items van de eendimensionele schaal gebruiken voor de totaalscore bij de Mantel-Haenszel-analyse dan blijken er 45 van de 360 geanalyseerde items partijdig te zijn (=13%). In totaal zijn er 13 items partijdig bij zowel de IRT- als de Mantel-Haenszel-procedure (=4%). Over het algemeen kan gesteld worden dat de IRT-procedure minder partijdige items opspoort dan de Mantel-Haenszel-techniek.
Items kunnen partijdig zijn in het voordeel of in het nadeel van Turkse en Marokkaanse leerlingen. Items blijken nooit in het voordeel te zijn voor Turkse leerlingen en tegelijkertijd in het nadeel van Marokkaanse leerlingen of omgekeerd.
Uit nadere analyse blijkt dat de beide procedures (met als totaalscore de items van de eendimensionele schaal) in 87% van de gevallen overeenstemmen in het detecteren van (on)partijdige items. Dit beeld komt overeen met de resultaten van andere onderzoekers (Bügel & Glas, 1991; Hambleton & Jones, 1992). De stabiliteit van de IRT- en de Mantel-Haenszel-procedure is vrijwel gelijk: in twee steekproeven wijst de Mantel-Haenszel-procedure bij 86% van de items en de IRT-procedure bij 89% van de items in beide steekproeven een item als partijdig aan.
Bij vergelijking van de resultaten van de verschillende Mantel-Haenszel-analyses blijkt dat het veel uitmaakt op welke items de totaalscore wordt gebaseerd. De analyses met de totaalscore gebaseerd op de 'gezuiverde' toets-onderdeelscore leveren in 37% van de gevallen partijdige items op, terwijl de analyses met de totaalscore op basis van de items van de eendimensionele schaal in 8% van de gevallen resulteren in partijdige items. Dit beeld wordt bevestigd door Clauser e.a. (1991).

Er is sprake van overlap tussen de verschillende procedures, maar er zijn ook verschillen waardoor het niet altijd duidelijk is of een item partijdig is of niet. Voordat gestart kan worden met de inhoudelijke analyse van partijdige items uit de Eindtoets Basisonderwijs 1987 en 1989 moet echter wel vastgesteld worden welke items partijdig zijn en welke niet. Hiermee komen we aan de vierde en vijfde onderzoeksvraag van deze dissertatie:

4 *Welke statistische procedure verdient de voorkeur voor het opsporen van itembias bij de Eindtoets Basisonderwijs?*

5 *Welke opgaven zijn voor allochtone leerlingen significant moeilijker of makkelijker dan voor autochtone leerlingen met een vergelijkbaar prestatieniveau?*

De verschillende resultaten worden met name beïnvloed door het feit dat de Mantel-Haenszel-procedure in tegenstelling tot de IRT-techniek gebaseerd is op de aanname dat het totaal aantal goed gemaakte opgaven een adequate schatting is van de te meten vaardigheid. Bij de IRT-procedure wordt deze assumptie getoetst, waardoor we er bij deze laatste procedure meer vanuit kunnen gaan dat we itembiasonderzoek doen met leerlingen van hetzelfde vaardigheidniveau. Omdat bij onderzoek naar itembias leerlingen juist op die vaardigheid gematcht moeten worden, gaat voor het detecteren van partijdige en onpartijdige items de voorkeur uit naar een itembiasdetectietechniek die gebaseerd is op het IRT-model. Hambleton & Jones (1992) en Dorans & Holland (1992) stellen dat IRT-procedures op theoretische gronden geprefereerd zouden moeten worden. De IRT-techniek biedt verder de mogelijkheid om niet-uniforme bias op te sporen (vgl. Hambleton & Rogers, 1989; Mellenbergh, 1989; Bügel & Glas, 1991).
Hoewel over het algemeen de voorkeur uitgaat naar IRT-procedures, stellen Hambleton & Rogers (1989), Hambleton & Jones (1992) en Dorans & Holland (1992) dat de Mantel-Haenszel-techniek als een goede vervanger beschouwd kan worden, hoewel erkend wordt dat de overeenstemming tussen beide procedures in het opsporen van partijdige items niet perfect is. Zij wijzen erop dat de betrouwbaarheid van de itembiasdetectieprocedures te wensen overlaat en dat moeilijk bepaald kan worden of een item nu wel of niet partijdig is. Sommige items zijn bij alle analyses partijdig, andere items zijn dit nooit en enkele items zijn bij een deel van de analyses partijdig.

Hoewel in theoretisch opzicht de voorkeur uitgaat naar IRT-procedures, verdient het feit, dat de Eindtoets Basisonderwijs samengesteld en geanalyseerd wordt volgens de klassieke testtheorie, nog nadrukkelijk aandacht. In het onderhavige onderzoek is wel gebleken dat de Eindtoetsitems passen op eendimensionele schalen, maar deze procedure is bij de psychometrische analyse en rapportage van de Eindtoets Basisonderwijs nog nooit gehanteerd. De scores van de deelnemers op de toetsonderdelen van de Eindtoets Basisonderwijs 1987 en 1989 zijn gebaseerd op het ongewogen aantal goed gemaakte opgaven. Verder is de totaalscore bepaald door de somscores van de drie toetsonderdelen op te tellen. Dit impliceert een arbitraire weging van het item en van de vaardigheidsdimensie: elk goed beantwoord item levert één punt op en elke schaal krijgt hetzelfde gewicht. Wanneer bij de rapportage van de Eindtoets Basisonderwijs 1987 en 1989 het eenparameter model zou zijn gehanteerd en weging van de schalen zou zijn toegepast, dan zou de totaalscore gebaseerd zijn op de gewogen somscore van items en schalen, hetgeen zowel voor allochtone als autochtone leerlingen tot andere toetsresultaten zou leiden. Omdat de correlaties tussen gewogen en ongewogen scores, respectievelijk schalen doorgaans hoog zijn te noemen (vgl. Bügel & Glas, 1991), zouden de totaalscores waarschijnlijk niet dramatisch verschillen van de in 1987 en 1989 aan de toetsdeelnemers gerapporteerde scores, maar het betekent wel dat we bij de keuze van de itembiasdetectieprocedure niet alleen vanuit theoretische voorkeuren kunnen vertrekken.
Gezien de bij de Eindtoets Basisonderwijs 1987 en 1989 gehanteerde scoringsprocedure ligt het voor de hand om een itembiasdetectieprocedure te kiezen die gebaseerd is op de klassieke testtheorie (vgl. Clauser e.a., 1991; Dorans & Holland, 1992; Schmitt e.a., 1992). Voor het antwoord op de vraag of een item

wel of niet partijdig is, wordt dan ook eerst gekeken naar de items die partijdig zijn volgens de Mantel-Haenszel-procedure met het 'gezuiverde' toetsonderdeel als totaalscore (zie 6.2.1). Deze items zijn vertrekpunt voor de uitgevoerde inhoudelijke analyse voor het opsporen van bronnen van itembias, zoals beschreven in hoofdstuk zeven.
De keuze voor de klassieke testtheorie-benadering heeft tot gevolg dat niet onderzocht wordt of de leerlingen inderdaad gematcht worden op dezelfde eendimensionele vaardigheid. Wanneer we meer in overeenstemming met de definitie van itembias (zie 1.2.2) in de Eindtoets Basisonderwijs 1987 en 1989 partijdige items willen detecteren, dan moet ook gekeken worden naar de items die met een IRT-procedure als zodanig zijn aangewezen (zie 6.2.2). Deze items zullen ook inhoudelijk worden geanalyseerd en er zal nagegaan worden of de resultaten van deze inhoudelijke analyses overeenstemmen met de resultaten van de eerder gemaakte inhoudelijke analyses. De mate van overeenstemming geeft een indicatie voor de graad van zekerheid waarmee we elementen van items als bron van itembias kunnen aanmerken (vgl. Clauser e.a., 1991; Dorans & Holland, 1992).

7 Bronnen van itembias

In het onderzoek naar itembias zijn twee elkaar aanvullende fasen onderscheiden. In de eerste fase zijn met statistische procedures partijdige items opgespoord. De eerste fase, de detectiefase, is beschreven in hoofdstuk zes. In de tweede fase wordt ingegaan op de vraag wat bij een bepaald item de oorzaak van itembias zou kunnen zijn. De tweede fase, de verklaringsfase, wordt in dit hoofdstuk beschreven.

Er is niet alleen in Nederland maar ook in andere landen weinig onderzoek gedaan naar oorzaken van itembias voor allochtone leerlingen. Volgens Schmitt, Holland & Dorans (1992) zijn er hiervoor drie redenen aan te wijzen. In de eerste plaats is onderzoek naar itembias relatief nieuw. Tot nu toe is de meeste aandacht uitgegaan naar statistische procedures voor het detecteren van partijdige items. In de tweede plaats veronderstelt het achterhalen van oorzaken van itembias voor allochtone leerlingen een theorie over de vraag waarom items voor de onderscheiden etnische groepen moeilijk zijn. Maar omdat de etnische groepen intern vaak in veel opzichten heterogeen zijn, kunnen de verschillen tussen de etnische groepen moeilijk beschreven worden. In de derde plaats is het opsporen van oorzaken van itembias complex, omdat bij een bepaald item verschillende oorzaken een rol kunnen spelen.

Omdat een theoretisch kader betreffende bronnen voor itembias voor allochtone leerlingen vooralsnog niet beschikbaar is, moeten we de conclusies die op basis van het onderhavige onderzoek in dit verband worden getrokken, beschouwen als voorlopig.

In 6.3 is aangegeven dat voor het antwoord op de vraag of een item uit de Eindtoets Basisonderwijs 1987 of 1989 wel of niet partijdig is, in eerste instantie wordt uitgegaan van de items die partijdig zijn volgens de Mantel-Haenszel-procedure met het 'gezuiverde' toetsonderdeel als totaalscore. Deze items worden eerst inhoudelijk geanalyseerd om mogelijke bronnen van itembias op te sporen. Daarna worden de items die volgens het IRT-model partijdig zijn eveneens aan een inhoudelijk analyse onderworpen. Vervolgens zal de overeenstemming tussen beide soorten analyses nagegaan worden.
Als vertrekpunt voor de inhoudelijke analyses zijn de items gekozen die volgens de Mantel-Haenszel-procedure met het 'gezuiverde' toetsonderdeel als totaal-score voor alle Turkse, respectievelijk Marokkaanse leerlingen partijdig zijn. Om echter ook de items met niet-uniforme itembias in de analyses te betrekken en om de stabiliteit, waarmee partijdige items worden gedetecteerd, te verhogen, zijn ook de items in de analyses betrokken die bij de Turkse en/of Marokkaanse leerlingen in drie of meer analyses partijdig zijn ($p<.01$). Dit betreft zowel items die partijdig zijn in het voordeel als die welke partijdig zijn in het nadeel van Turkse of Marokkaanse leerlingen. Het aantal items dat partijdig is in het voordeel van Turkse en Marokkaanse leerlingen is gering. Omdat de verschillen tussen bias in het voordeel of in het nadeel meer aanwijzingen over bronnen van itembias kunnen geven, zijn ook items inhoudelijk onderzocht die bij alle vijf analyses negatieve z-waarden hebben.

Aan het zoeken naar mogelijke oorzaken van itembias hebben niet alleen de medewerkers van het onderzoeksproject (van KUB en Cito), maar ook niet bij het project betrokken experts en leerlingen uit groep acht van het basisonderwijs een bijdrage geleverd.
De medewerkers van het onderzoeksproject, drie taalkundigen (KUB) en een onderwijskundige (Cito), hebben onafhankelijk van elkaar, op grond van hetgeen de items beogen te meten, partijdige items geanalyseerd. Bij deze inhoudsanalyse hebben de in 2.2 geformuleerde mogelijke bronnen van itembias voor allochtone leerlingen als hypothesen gefungeerd. De gemeenschappelijkheden in de analyses van de projectmedewerkers zijn geïnventariseerd en die worden beschreven in 7.1.
Om meer aanwijzingen over bronnen van itembias te verkrijgen zijn ook niet bij het project betrokken experts gevraagd partijdige items op mogelijke biasbronnen te beoordelen. In 7.2 wordt ingegaan op de oordelen van experts over mogelijke bronnen van itembias. De medewerkers van het onderzoeksproject hebben ook leerlingen uit groep acht van het basisonderwijs in het onderzoek betrokken. Door middel van een kleinschalig hardop-denken-experiment is nagegaan hoe vaak allochtone en autochtone leerlingen partijdige items fout beantwoorden door een itemelement dat vermoedelijk de bron van itembias is. Daarnaast is onderzocht hoe vaak bij gemanipuleerde items door de itemmanipulatie een goed antwoord wordt gegeven. Gemanipuleerde items zijn items waarbij het itemelement dat als potentiële biasbron is aangewezen, vervangen is door een itemelement waarvan verwacht wordt dat het geen bias veroorzaakt. In 7.3 wordt verslag gedaan van het hardop-denken-experiment waarin leerlingen uit groep acht van het basisonderwijs aangeven hoe ze de oorspronkelijke (partijdige), respectievelijk de gemanipuleerde items hebben opgelost.

7.1 Inhoudelijke analyse van partijdige items

De vier projectmedewerkers hebben onafhankelijk van elkaar de items inhoudelijk geanalyseerd met het oog op de vraag welke itemelementen mogelijk de bron van itembias vormen. Hierbij ging speciale aandacht uit naar itemelementen die voorkomen in items die partijdig zijn in het nadeel van beide groepen leerlingen en ontbreken in items die in het voordeel zijn van deze leerlingen en omgekeerd. Bij deze inhoudsanalyse hebben de potentiële bronnen van itembias voor allochtone leerlingen uit 2.2 gefungeerd als hypothesen. Uit de gemeenschappelijkheden in de analyses van de projectmedewerkers bleek dat de inhoudelijke analyses twee fundamentele problemen opleveren.

7.1.1 Problemen bij de inhoudelijke analyse van partijdige items

a Bij de inhoudelijke analyse van de partijdige items bleek dat het uitermate moeilijk is om met zekerheid aan te geven welk itemelement nu precies de bron van itembias vormt. Als voorbeeld bespreken we hieronder de rekenitems 49 en 59 uit 1989.

1989 Rekenen nr. 49 (taak Rekenen 2 nr. 19)

$\frac{3}{5}$ deel van de weg van Strandoord naar Bosdorp is geasfalteerd.
De rest (30 km) moet nog gedaan worden.
Hoelang is de weg van Strandoord naar Bosdorp?

A 12 km
B 45 km
C 50 km
D 75 km

Dit item is in statistisch opzicht partijdig in het nadeel van de Turkse leerlingen en onpartijdig voor Marokkaanse leerlingen. Waar is nu in dit item voor Turkse leerlingen de bron van bias gelegen? Het is mogelijk dat de biasbron wordt gevormd door sommige talige elementen van het item. De woordenschat van de leerling moet het mogelijk maken dat de leerling begrijpt dat een bepaald iets voor drievijfde deel af is en dat dat 'af-zijn' geassocieerd moet worden met 'geasfalteerd'. Het moet ook duidelijk zijn dat 'De rest (30 km)' een verwijzing is naar het tweevijfde deel dat nog niet af is en dat dat gelijk is aan 30 kilometer. Vervolgens moet begrepen worden dat met 'Hoelang is de weg van Strandoord naar Bosdorp?' gevraagd wordt het geheel (vijfvijfde deel) uit te rekenen. De biasbron kan ook nog in het woord 'Hoelang' gelegen zijn. Het is mogelijk dat 'Hoelang' voor verwarring zorgt, omdat het woord betrekking kan hebben op tijd en ook op hoeveelheid.
De bron van itembias kan daarnaast ook in de rekenvaardigheid zitten die nodig is om de rekenbewerking te verrichten: de rest vormt het tweevijfde deel = 30 km; éénvijfde deel = 15 km; het geheel = 75 km.
Het is denkbaar om de bron van itembias in eerste instantie in talige elementen te zoeken. Maar welk tekstelement geeft Turkse leerlingen onvoldoende duidelijkheid over de te verrichten rekenoperatie? Voor het juist beantwoorden van het volgende item is ook veel taalvaardigheid Nederlands vereist, maar opmerkelijk is dat dit item – hoewel niet significant – bij alle analyses in het voordeel is van Turkse leerlingen. Wellicht geeft de tekst bij dit item Turkse leerlingen voldoende aanknopingspunten om de opgave juist op te lossen, terwijl de formulering van de vraag met het werkwoord 'zich verhouden tot' en het verwijswoord 'dat' toch niet eenvoudig is. Het is mogelijk dat de leerlingen door het rekenonderwijs op de basisschool vertrouwd zijn geworden met een idiomatische uitdrukking als 'zich verhouden tot'.

1989 Rekenen nr. 59 (taak Rekenen 2 nr. 29)

Dolf heeft *f* 3,– meer dan Michel. Michel heeft *f* 5,–.
Hoe verhoudt zich Dolfs geld tot dat van Michel?

A 2 : 5
B 3 : 5
C 5 : 3
D 8 : 5

Omdat rekenopgave 59 uit 1989 bij alle analyses – niet significant – in het voordeel is van Turkse leerlingen, hebben deze leerlingen kennelijk voldoende rekenvaardigheid om deze operationalisatie van het rekendomein Verhoudingen juist te beantwoorden. Dit is opmerkelijk omdat in 4.2 bleek dat er aanwijzingen zijn dat Turkse leerlingen over het algemeen opgaven over verhoudingen moeilijk vinden. Het rekenonderdeel verhoudingen komt traditioneel relatief laat in het basisonderwijs aan de orde en scholen met meer traditionele rekenmethoden (hieronder zijn de scholen met relatief veel allochtone leerlingen oververtegenwoordigd) besteden minder tijd aan de instructie over verhoudingen dan scholen met modernere methoden (Wijnstra, 1988). Het bovenstaande geeft aanleiding te veronderstellen dat de rekenvaardigheid bij opgaven over verhoudingen naast de taalvaardigheid een bron van bias kan zijn, maar opgave 59 uit 1989 is juist – niet significant – in het voordeel van Turkse leerlingen.
De beide voorbeelden geven aan dat een item een aantal talige elementen bevat die elk afzonderlijk of in combinatie met elkaar biasbron kunnen zijn, evenals elke te verrichten rekenoperatie. De beide rekenitems doen door hun talige context een beroep op taalvaardigheid Nederlands.
De contexten die bij items gebruikt worden, variëren aanzienlijk, ook in complexiteit. Zo wordt bij het toetsonderdeel Informatieverwerking als contextmateriaal zowel van grafische afbeeldingen als van teksten gebruik gemaakt. Ook de voorkennis over de onderwerpen die in het contextmateriaal aan de orde worden gesteld (bijvoorbeeld: voorkennis over het onderwerp van een tekst), kunnen bron van bias zijn. Kortom: het blijkt moeilijk om daadwerkelijk de biasbron te isoleren.

b De inhoudelijke analyse van de items wordt ook bemoeilijkt door het feit dat bij inhoudelijk sterk vergelijkbare items de ene keer wel en de andere keer niet sprake is van partijdigheid. Als voorbeeld bespreken we hieronder de items Rekenen 1987, nummer 27 versus Rekenen 1989, nummer 26 en Rekenen 1987, nummer 41 versus Rekenen 1989, nummer 40.

1987 Rekenen nr. 27 (taak Rekenen 1 nr. 27)

Welke van de volgende zinnen past hier het best?

A In dit bos staan 75 beuken.
B 3 van de 4 bomen in dit bos zijn beuken.
C $\frac{1}{75}$ deel van het bos bestaat uit beuken.
D Van elke 75 bomen is in dit bos gemiddeld 1 boom een beuk.

1989 Rekenen nr. 26 (taak Rekenen 1 nr. 26)

Wat betekent dit?

A 1 van elke 60 mensen doet te weinig aan sport.
B 60 mensen doen te weinig aan sport.
C $\frac{1}{60}$ deel van de mensen doet te weinig aan sport.
D Het goede antwoord staat er niet bij.

Item 27 uit 1987 is in sterke mate partijdig in het nadeel van zowel Turkse als Marokkaanse leerlingen. Item 26 uit 1989 is – hoewel niet significant – in het voordeel van Marokkaanse leerlingen. De items gaan inhoudelijk gezien over dezelfde rekendoelstelling: omzetten van percentages in verhoudingsgetallen (Cito, 1986a: 33). De geringe verschillen tussen beide operationalisaties van de doelstelling geven geen aanleiding te veronderstellen dat de rekenvaardigheid als zodanig de ene keer wel en de andere keer niet biasbron is. Het grootste verschil tussen beide items is wellicht gelegen in het juiste antwoord: B, respectievelijk D. '3 van de 4 bomen' is als uitdrukking weinig frequent en verder komt 'Het goede antwoord staat er niet bij' als goede antwoord in de toets niet vaak voor. Aan de andere kant kan gesteld worden dat noties als '3 van de 4' in het rekenonderwijs veelvuldig gehanteerd worden en derhalve als onderdeel van de schoolse 'rekentaal' te beschouwen zijn. Ook moet opgemerkt worden dat leerlingen die op school met het oog op de toetsafname vertrouwd zijn gemaakt met de verschillende itemtypen van de Eindtoets Basisonderwijs (bijvoorbeeld door het maken van een Eindtoets van een voorafgaand jaar) over het algemeen de functie van alternatief D bij rekenopgave 26 uit 1989 zullen herkennen. Het is mogelijk dat talige elementen item 27 uit 1987 moeilijker maken dan item 26 uit 1989. Zo is de vraag 'Wat betekent dit?' directer dan de vraag 'Welke van de volgende zinnen past hier het best?'. Toch is het moeilijk om in te zien dat de genoemde verschillen tussen beide items kunnen verklaren dat item 27 uit 1987 sterk partijdig is in het nadeel van Turkse en Marokkaanse leerlingen en dat item 26 uit 1989 – niet significant – in het voordeel van Marokkaanse leerlingen is.

De volgende twee items lijken inhoudelijk eveneens op elkaar. Beide zijn ze operationalisaties van de rekendoelstelling 'Berekenen van het gemiddelde' (Cito, 1986a: 26)

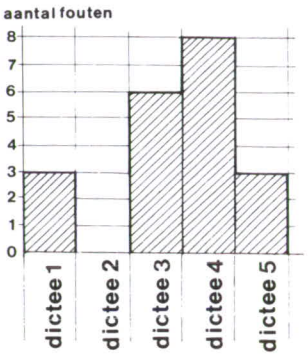
1987 Rekenen nr. 41 (taak Rekenen 2 nr. 11)

Wat is het gemiddelde van de volgende getallen:

1 ; 2 ; 3 ; 99 ; 98 ; 97

A 50
B 100
C 150
D 300

1989 Rekenen nr. 40 (taak Rekenen 2 nr. 10)



Nicolien houdt met een grafiekje bij hoeveel fouten zij maakt in haar dictees.
Wat is het gemiddelde aantal fouten per dictee?

A 3
B 4
C 5
D 6

Rekenitem 41 uit 1987 is in sterke mate partijdig in het nadeel van zowel Turkse als Marokkaanse leerlingen, terwijl item 40 uit 1989 slechts bij twee analyses partijdig is in het nadeel van Turkse en bij alle analyses onpartijdig is voor Marokkaanse leerlingen. Inhoudelijk gezien lijkt item 40 uit 1989 echter meer potentiële bronnen van itembias te bevatten dan item 41 uit 1987. Item 40 bevat immers meer talige en grafische elementen. Bovendien moet de leerling bij dit item uitzoeken van welke getallen het gemiddelde berekend moet worden en daarbij moet de leerling opmerken dat dictee 2 foutloos is gemaakt. Voor beide items geldt dat waarschijnlijk alle leerlingen over de benodigde woordenschat beschikken, omdat de gehanteerde begrippen in het onderwijsprogramma van vrijwel alle basisscholen voorkomen. De te verrichten rekenkundige bewerking lijkt bij item 41 complexer, maar door het handig combineren van getallen is ook dit item uit het hoofd uit te rekenen.
Items die inhoudelijk sterk vergelijkbaar zijn, vormen een belangrijk object voor inhoudelijke analyses. Hetgeen de items beogen te meten is voor deze items gelijk of sterk vergelijkbaar en heeft bij beide items in principe een even grote kans biasbron te zijn.

Uit de gepresenteerde voorbeelditems komt naar voren dat het soms moeilijk is om met zekerheid aan te geven waarom het ene item in sterkere mate partijdig is dan het andere. Toch moet aan de andere kant vastgesteld worden dat er een aanzienlijk aantal partijdige items is met opvallende onderlinge overeenkomsten. Deze items geven sterke aanwijzingen voor inhoudelijke bronnen van itembias. In 7.1.2 wordt hierop nader ingegaan en worden de belangrijkste resultaten van de uitgevoerde inhoudelijke biasanalyses gegeven.

7.1.2 Eerste resultaten van de inhoudelijke analyse van partijdige items

Zoals eerder is vermeld, hebben de medewerkers van het onderzoeksproject, drie taalkundigen (KUB) en een onderwijskundige (Cito), onafhankelijk van elkaar de in statistisch opzicht partijdige items in het voor- of nadeel van Turkse en/of Marokkaanse leerlingen inhoudelijk geanalyseerd. Bij deze inhoudsanalyse hebben de in 2.2 geformuleerde potentiële bronnen van itembias voor allochtone leerlingen gefungeerd als hypothesen. Vervolgens zijn de gemeenschappelijkheden in de analyses van de projectmedewerkers geïnventariseerd. Hieruit is gebleken dat er over een groot aantal items overeenstemming bestaat inzake mogelijke bronnen van itembias. Uit de analyses bleek dat op grond van de in de items in het geding zijnde vaardigheden itemclusters gevormd kunnen worden. Items behorende bij een itemcluster vertonen overeenkomsten ten aanzien van hetgeen de items beogen te meten of vertonen overeenkomsten met betrekking tot additionele vaardigheden die relevant zijn voor het juist beantwoorden van de items. De inhoudelijke analyses leveren voorlopige conclusies op (vgl. Uiterwijk & Vallen, 1991). De conclusies worden voorlopig genoemd, omdat de conclusies in een volgende fase van het onderzoek zijn vergeleken met de oordelen van niet bij het onderzoeksproject betrokken experts (zie 7.2).
Bij deze analyses zijn die items betrokken die bij deze leerlingen in drie of meer van de in 6.2.1 besproken analyses partijdig zijn ($p<.01$). Om meer aanwijzingen over bronnen van itembias te krijgen, zijn echter ook items inhoudelijk onderzocht, die bij alle vijf uitgevoerde analyses – niet significante – negatieve z-waarden hebben.
Voordat we beginnen met de bespreking van de resultaten van de inhoudelijke analyses per itemcluster, geven we in tabel 7.1 een overzicht van de onderscheiden itemclusters met het aantal items, dat volgens de Mantel-Haenszel- of de IRT-procedure partijdig is in het voor- of nadeel van Turkse en/of Marokkaanse leerlingen.

Tabel 7.1 Overzicht van de itemclusters met het aantal partijdige items per cluster

Cluster en toetsonderdeel		Bias in nadeel			Bias in voordeel		
		Tu	Ma	Tu & Ma	Tu	Ma	Tu & Ma
A	Tekststructuur (T)	1	1	6	1		
			1				
B	Tekstbegrip (I)	6		9	3	1	
			1	**2**			**2**
C	Woordkennis en kennis	1	1	7			
	van woordcombinaties (T)	**2**	**1**	**3**		**1**	
D	Figuurlijk taalgebruik (T & I)	3		1			
E	Correct taalgebruik (T)	2	1	10	1	1	1
		1		**1**			**1**
F	Referenties (T & I)	3		4			
G	Spelling (T)	2			8	3	2
					3		
H	Keuze en gebruik van informatiebronnen (I)		2	2	1		
I	Kennen en kunnen gebruiken van rekenkundige begrippen (R)			2			
J	Herleidingen (R)	1	1				
K	Omzetten van verhoudingen in procenten en omgekeerd (R)	1		2			
L	Verhoudingen (R)	2		2			
M	Rekenitems met veel context (R)	4	1	2			
	Rekenitems met weinig of geen context (R)			1	1	4	
			1				
N	Items met grafische context (I)		6	3	1		

Toelichting:
Een niet-vetgedrukt cijfer geeft het aantal items aan dat partijdig is volgens de Mantel-Haenszel-procedure; een vetgedrukt cijfer geeft het aantal items aan dat partijdig is volgens de IRT-procedure
T = Taal, R = Rekenen, I = Informatieverwerking
Tu = Turkse leerlingen, Ma = Marokkaanse leerlingen

A Tekststructuur (toetsonderdeel Taal)

Van de volgens de Mantel-Haenszel-analyses partijdige taalitems die betrekking hebben op de tekststructuur zijn er
– zes in het nadeel van Turkse en Marokkaanse leerlingen,
– één in het nadeel van Marokkaanse leerlingen,
– één in het nadeel en
– één in het voordeel van Turkse leerlingen.
In het toetsonderdeel Taal van de Eindtoets Basisonderwijs (Taal wordt met een hoofdletter geschreven, wanneer verwezen wordt naar het taalonderdeel van de Eindtoets Basisonderwijs) zijn teksten opgenomen die taalkundig gezien tekorten vertonen en de items bij de tekst gaan na of leerlingen in staat zijn de

gebreken in de tekst te herstellen. De leerlingen moeten hierbij meestal zinnen herstellen, verplaatsen of invoegen. Items hebben een grote kans partijdig te zijn in het nadeel van Turkse en Marokkaanse leerlingen wanneer de in te voegen of te verplaatsen zin een verwijzingselement bevat en wanneer bij de oplossing niet volstaan kan worden met een globaal begrip van de totale tekst of van de betreffende grotere passage. Partijdigheid treedt derhalve vooral op wanneer een nauwkeurig begrip van een woord of van één of enkele zinnen noodzakelijk is. Dit sluit aan bij de onderzoeksresultaten van Hacquebord (1989). Zij constateerde dat voor Turkse leerlingen tekstbegripitems op het microniveau van een tekst (woord- en zinsniveau) moeilijker zijn, dan tekstbegripitems op het meso- (alinea-niveau) en op het macroniveau (het hoofdthema, de tekstsoort, de strekking) van een tekst.
Items in het cluster Tekststructuur die betrekking hebben op het weghalen van redundanties in een tekst of op het verwijderen van fouten in de opbouw van een tekst geven eveneens een grote kans op itembias in het nadeel van deze leerlingen. De volgende twee items over het verplaatsen, respectievelijk invoegen van een zin zijn in hoge mate partijdig in het nadeel van Turkse en Marokkaanse leerlingen.

1987 Taal nr. 33 (taak Taal 2 nr. 13)

1 Zoals jullie weten, zijn de mensen van de reddingsbrigade vaak dapper. Zij geven
2 zich vrijwillig op voor dit werk. Soms riskeren ze zelfs hun leven bij het redden van
3 mensen in nood. Vorige week zaterdag moest de reddingsbrigade alweer een surfer
4 helpen. Die was tegen beter weten in te ver de zee opgegaan. Er stond die dag een
5 hele harde wind. Daardoor was de surfer erg vermoeid geraakt. Hij kon toen niet
6 meer op zijn plank komen om zijn zeil omhoog te trekken. Door de sterke wind en de
7 zeestroming dreef hij steeds verder van het strand af. Leden van de reddingsbrigade
8 hadden hem niet te laat met hun verrekijkers opgemerkt. Ze gingen er met hun snelle
9 speedboot meteen op af. Ze hadden geen vijf minuten later moeten komen. Twintig
10 minuten later bracht de bemanning de uitgeputte surfer weer veilig aan land bij de
11 uitkijkpost. Zonder die reddingsbrigade zouden er, denk ik, heel wat onverstandige
12 mensen verdrinken. Volgend jaar wil ik ook graag bij de reddingsbrigade komen.
13 Lijkt jullie dat ook niet leuk?

Vorige week zaterdag moest de reddingsbrigade alweer een surfer helpen. (r. 3, 4)
Wat kun je het beste doen met deze zin?

A Zo laten staan.
B Plaatsen achter: ... werk. (r. 2)
C Plaatsen achter: ... geraakt. (r. 5)
D Plaatsen achter: ... trekken. (r. 6)

1987 Taal nr. 34 (taak Taal 2 nr. 14)

(zie de tekst bij item 1987 Taal nr. 33; taak Taal 2 nr. 13)

Anders zou hij vast en zeker zijn verdronken.
Waar kun je deze zin het beste plaatsen?

A Achter: ... opgegaan. (r. 4)
B Achter: ... af. (r. 7)
C Achter: ... komen. (r. 9)
D Achter: ... verdrinken. (r. 12)

De items 33 en 34 uit 1987 vragen telkens bepaalde passages uit de tekst nauwkeurig te lezen en de betekenis ervan in relatie met de gegeven zin te beoordelen. Globaal begrip van de totale tekst is eveneens van belang (de leerlingen zijn geïnstrueerd om voor het beantwoorden van de items eerst de gehele tekst te lezen). De te verplaatsen en de in te voegen zin bevatten ingewikkelde verwijzingselementen.
Het verwijzingselement 'alweer' in taalitem 33 uit 1987 kan tot problemen leiden, omdat met 'alweer' in feite naar situaties (voorgaande surfers in nood) wordt verwezen die niet in de tekst voorkomen.
Het woord 'Anders' in de in te voegen zin van taalitem 34 uit 1987 legt een verbinding met de voorgaande zin. Het leggen van het verband met de voorgaande zin wordt bemoeilijkt, omdat de leerling eerst moet uitzoeken waar de zin het beste geplaatst kan worden. De verwijzingselementen in deze items zijn van belang omdat ze aanwijzingen geven voor het juist oplossen van het item.

Van de items die volgens de IRT-analyses partijdig zijn, behoort item 1989 Taal nummer 31 eveneens tot het cluster Tekststructuur. Dit item is alleen partijdig in het nadeel van Marokkaanse leerlingen. Het heeft betrekking op het opsporen van ontbrekende elementen in een tekst gezien de functie van die tekst (in een brief ontbreekt de datum van een bijeenkomst, waarvoor de briefschrijver de lezer uitnodigt). Dit item geeft voeding aan de veronderstelling dat items die betrekking hebben op Tekststructuur een bron kunnen zijn van itembias, maar het betreft slechts één item en één kenmerk van het cluster Tekststructuur.

B Tekstbegrip (Informatieverwerking)

Van de partijdige items uit het onderdeel Informatieverwerking die betrekking hebben op tekstbegrip zijn er
- negen in het nadeel van Turkse en Marokkaanse leerlingen,
- zes items zijn in het nadeel van Turkse leerlingen,
- drie items zijn in het voordeel van Turkse leerlingen en
- één item is in het voordeel van Marokkaanse leerlingen.

Twee items die partijdig zijn in het voordeel van Turkse leerlingen, hebben betrekking op globaal tekstbegrip (bijvoorbeeld: Waarover gaat het in dit stukje

tekst vooral?). Onder de items die bij alle analyses – niet significant – in het voordeel van Turkse en/of Marokkaanse leerlingen zijn, komen ook relatief veel items voor die naar de hoofdgedachte van een (deel van een) tekst vragen. Daar staat echter tegenover dat twee partijdige items in het nadeel van beide groepen leerlingen en één item in het nadeel van Turkse leerlingen op dezelfde wijze eveneens naar globaal tekstbegrip vragen.
Dat bij globaal tekstbegrip de kans op itembias in het nadeel van deze leerlingen niet groot is, wordt ondersteund door het feit dat het grootste deel van de overige items, die partijdig zijn in het nadeel van beide groepen leerlingen, meer betrekking hebben op het meso- (alinea-niveau) en micro-niveau (woord- en zinsniveau) van de tekst (vgl. Hacquebord, 1989). Deze items vragen veelal om een woordelijke of geparafraseerde herhaling van expliciet in de tekst gegeven informatie. Als voorbeeld hiervan geldt 1987 Informatie-verwerking nr. 4.

1987 Informatieverwerking nr. 4 (taak Informatieverwerking 1 nr. 4)

Een opinie-onderzoek

1 Bijna tweederde van de Britse ouders vindt het goed dat hun kind stokslagen op
2 school krijgt als straf voor wangedrag. Uit een opinie-onderzoek, gepubliceerd in
3 dagblad The Times, blijkt dat 65 procent van de ondervraagden toestemming zou
4 geven voor het leggen van de stok over de billen van hun kinderen. Drieëndertig
5 procent van de ouders is tégen lijfstraf .....

Hoe denken de Engelse ouders volgens het opinie-onderzoek over lijfstraffen op school (r. 1 t/m 5)?

A Een meerderheid is voor lijfstraffen bij wangedrag.
B Een meerderheid is van mening dat alleen ouders over lijfstraffen mogen beslissen.
C Een meerderheid vindt dat er vaker lijfstraffen op scholen gegeven moeten worden.
D Een meerderheid vindt dat lijfstraffen vanaf 1949 niet hadden mogen toenemen.

De constatering dat bij items die betrekking hebben op globaal tekstbegrip (macro-niveau) de kans op itembias in het nadeel van deze leerlingen niet groot is, sluit aan bij de bevindingen van Hacquebord (1989). Zij vond dat 10% van de variantie in tekstbegripscores op microniveau door de etnische achtergrond van de leerlingen wordt verklaard, terwijl deze achtergrond 2% van de variantie in de scores op mesoniveau verklaart en 1% op macroniveau (zie 2.2.2).
Verhoeven & Vermeer (1992) wijzen in dit verband met name op het belang van woordkennis bij het lezen van teksten door allochtone leerlingen. Hacquebord (1989) is van mening dat tweetalige leerlingen hun geringe taal-kennis op het microniveau van een tekst compenseren met efficiënte lees-strategieën op het macroniveau van een tekst. Woordkennis op zich is weliswaar een belangrijke, maar geenszins voldoende voorwaarde voor de complexe tekstbegripvaardigheid (Hacquebord, 1989: 253).
Als we de items die bij een bepaalde tekst horen bezien, dan zijn er geen aanwijzingen dat de ene tekst aanzienlijk meer of minder partijdige items bevat

dan de andere. Bij vrijwel alle teksten zijn partijdige en onpartijdige items en er is geen duidelijk verband te constateren met het onderwerp van de tekst.
De vijf items die volgens de IRT-analyses partijdig zijn bevestigen over het algemeen het zojuist geschetste beeld. De items waarbij de betekenis van een woord centraal staat, zijn bijna alle partijdig in het nadeel van Turkse en Marokkaanse leerlingen (bijvoorbeeld: Met het woord *gretig* wordt uitgedrukt dat ...). Bij de items die partijdig zijn in het voordeel moeten de leerlingen aangeven wat de referent is van een bepaald woord (bijvoorbeeld: Waarnaar verwijst *die*?). Er moet echter worden bedacht dat in deze items het verwijswoord meestal wel relatief dicht in de buurt van de referent staat en een aanduiding is voor een concreet object (de zee, een kievit).

C Woordkennis en kennis van woordcombinaties (toetsonderdeel Taal)

Van de partijdige items uit het toetsonderdeel Taal die betrekking hebben op de kennis van woorden of woordcombinaties zijn er
- zeven in het nadeel van Turkse en Marokkaanse leerlingen,
- één item is in het nadeel van Turkse leerlingen en
- één item is in het nadeel van Marokkaanse leerlingen.

Er komen in de Eindtoets Basisonderwijs 1987 en 1989 geen items over woordkennis en kennis van woordcombinaties voor die partijdig zijn in het voordeel van beide groepen leerlingen. De negen tot dit cluster behorende items zijn te gering in aantal om specifieke categorieën woorden (bijvoorbeeld: voorzetsels, bijwoorden) als bron van itembias aan te wijzen. Over het algemeen zijn er wel aanwijzingen dat moeilijke woorden waarvan de betekenis niet of moeilijk uit de context kan worden afgeleid een grotere kans hebben om bron van itembias te zijn dan moeilijke woorden waarvan de betekenis wel uit de context kan worden afgeleid. Als voorbeeld geldt 1989 Taal nr. 34.

1989 Taal nr. 34 (taak Taal 2 nr. 14)

43 Ik hoop echt dat jullie komen. Jullie moeten wel uiterlijk half zeven bij ons thuis zijn.
44 Jullie moeten echt zo vroeg mogelijk komen hoor. Anders zijn er vast geen kaartjes
45 meer te krijgen.

Wat had Saskia in plaats van *uiterlijk* (r. 43) ook kunnen schrijven?

A op zijn best
B op zijn laatst
C op zijn mooist
D op zijn vroegst

In taalitem 34 uit 1989 moeten de leerlingen de betekenis van het woord *uiterlijk* aangeven. De tekst waar het item betrekking op heeft, geeft in feite geen steun voor de betekenis van het woord. In 2.2.2 is reeds de ondersteunende rol genoemd, die de context kan hebben bij het aangeven van de betekenis van een woord (vgl. Tabossi, 1991). Item 34 uit 1989 is te beschouwen als een voorbeeld van partijdige items die voeding geven aan de

veronderstelling dat moeilijke woorden waarvan de betekenis niet of moeilijk uit de context kan worden afgeleid een grotere kans hebben om bron van itembias te zijn. Het is niet verwonderlijk dat er in een toets waarin naar woordkennis gevraagd wordt, relatief veel items voorkomen waarbij de context weinig aanknopingspunten biedt voor het aangeven van de betekenis van het woord. Er wordt immers afbreuk gedaan aan de constructvaliditeit van de toets wanneer de items over woordkennis ook door andere vaardigheden goed kunnen worden beantwoord. In dit verband kan met betrekking tot andere vaardigheden gedacht worden aan 'het goed kunnen lezen en begrijpen van de passages voor en na het betreffende woord'.

Van de items die volgens de IRT-analyses partijdig zijn, behoren zeven items tot het cluster Woordkennis en kennis van woordcombinaties. Deze items maken, op één na, alle deel uit van de reeds in dit cluster besproken items. Bijna alle items bevestigen de veronderstelling dat moeilijke woorden waarvan de betekenis niet of nauwelijks uit de context kan worden afgeleid, een grotere kans maken om als biasbron te fungeren.

D Figuurlijk taalgebruik (toetsonderdelen Taal en Informatieverwerking)

Van de partijdige items die betrekking hebben op de kennis van de betekenis van figuurlijk taalgebruik is er
- één in het nadeel van Turkse en Marokkaanse leerlingen en zijn er
- drie in het nadeel van Turkse leerlingen.

De items hebben betrekking op weinig gebruikte idiomatische uitdrukkingen, gezegden en beeldspraken. Bij het item dat partijdig is in het nadeel van beide groepen leerlingen moet uit vier gezegden het gezegde gekozen worden dat het beste past bij een bepaalde uitspraak (bijvoorbeeld: Het is niet alles goud wat er blinkt). Cacciari & Glucksberg (1991) vonden dat de betekenis van figuurlijk taalgebruik moeilijker te geven is, wanneer de afstand tussen de figuurlijke en letterlijke betekenis groter is. Zij wijzen er tevens op dat ook van belang is in welke mate de context de interpretatie van het spreekwoord of gezegde ondersteunt.
In dit cluster bevinden zich twee items die – niet significant – in het voordeel zijn van Marokkaanse leerlingen. Het volgende item is partijdig in het nadeel van Turkse kinderen en – niet significant – in het voordeel van Marokkaanse leerlingen.

1987 Taal nr. 2 (taak Taal 1 nr. 2)

Welk bijvoeglijk naamwoord is figuurlijk gebruikt?

A een gouden ketting
B een koperen kraan
C een zilveren bruiloft
D een zinken boot

Gezien het geringe aantal partijdige items in dit cluster kan vooralsnog niet met enige stelligheid gezegd worden dat figuurlijk taalgebruik bij Turkse leerlingen een grotere kans op bias heeft dan bij Marokkaanse leerlingen; de items geven wel aanwijzingen in deze richting.
Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster Figuurlijk taalgebruik.

E Correct taalgebruik (toetsonderdeel Taal)

Van de partijdige items die betrekking hebben op Correct taalgebruik zijn er
- tien in het nadeel van zowel Turkse als Marokkaanse leerlingen,
- twee items zijn in het nadeel van Turkse en
- één item is in het nadeel van Marokkaanse leerlingen,
- één item is in het voordeel van beide groepen leerlingen,
- één item is in het voordeel van Turkse en
- één item is in het voordeel van Marokkaanse leerlingen.

De items in dit cluster vragen van de leerlingen om woorden en zinnen in teksten qua vorm op hun correctheid te beoordelen en zonodig aan te geven op welke wijze een woord of zin het beste verbeterd kan worden. De items die partijdig zijn in het nadeel van beide groepen leerlingen hebben in grote mate betrekking op kennis van de vorm van vaste woordcombinaties en conventies op het gebied van de zinsbouw. Voor het goed beantwoorden van deze items speelt over het algemeen de betekenis van de woordcombinatie niet of nauwelijks een rol. Item 1987 Taal nr. 26 heeft betrekking op de kennis van de vorm van een vaste woordcombinatie en is partijdig in het nadeel van beide groepen leerlingen.

1987 Taal nr. 26 (taak Taal 2 nr. 6)

30 Bijna gelijktijdig schieten Kokkie en Appel de ruimte in. Ze komen aan de voet van
31 een zandberg neer. De bemanning is er zonder kleerscheuren afgelopen. Alleen
32 Appel heeft een paar blauwe plekken opgelopen.

Wat kun je het beste doen met: *afgelopen.* (r. 31)?

A Zo laten staan
B Vervangen door: afgekomen
C Vervangen door: afgeraakt
D Vervangen door: afgevallen

Het item dat partijdig is in het voordeel van Turkse en Marokkaanse leerlingen (1989 Taal nr. 16) heeft betrekking op het gebruik van de juiste tijd van het werkwoord in een zin, gezien de tijdsreferentie in voorgaande en volgende zinnen.

1989 Taal nr. 16 (taak Taal 1 nr. 16)

23 Ja, het lijkt me heerlijk om in die oude denneboom te wonen.Alleen weet ik niet of
24 het 's winters niet te koud zou zijn. Ik denk ook dat mijn broertje me erg zou missen.
25 En mijn vader die altijd zegt: "Als ik jou niet had ...." Wat zou het naar voor hem
26 zijn. Toch leek het me wel leuk om de hele dag lekker te doen waar ik zin in heb. Ik
27 zal dan niet meer horen: "Je moet eten, je moet naar tennissen, je moet je kamer
28 opruimen, en je moet naar bed."

Wat kun je het beste doen met: *Toch leek het me wel leuk* ... (r. 26)

A Zo laten staan
B Vervangen door: Toch had het me wel leuk geleken ...
C Vervangen door: Toch heeft het me wel leuk geleken ...
D Vervangen door: Toch lijkt het me wel leuk ...

Het item dat partijdig is in het voordeel van Marokkaanse leerlingen heeft betrekking op de vorm van een vaste woordcombinatie (het opvoeren van een musical). Omdat er ook items over het gebruik van de juiste tijd van het werkwoord en items over de vorm van vaste woordcombinaties zijn die partijdig zijn in het nadeel, bieden de partijdige items in het voordeel weinig aanknopingspunten voor bronnen van itembias. De in totaal 13 partijdige items in het nadeel van Turkse en Marokkaanse leerlingen geven aanleiding te veronderstellen dat kennis van de vorm van woordcombinaties en van conventies op het gebied van de zinsbouw bron van itembias kunnen zijn. Dit sluit aan bij Kerkhoff (1988) die vond dat onder andere Turkse en Marokkaanse leerlingen meer syntactische fouten maakten dan autochtone leerlingen. In haar onderzoek zijn de meest voorkomende syntactische fouten het niet gebruiken van één of meer noodzakelijk vereiste woorden in een zin en het gebruiken van verkeerde voegwoorden (zie 2.2.2).

Van de items die volgens de IRT-analyses partijdig zijn, behoren er drie items tot het cluster Correct taalgebruik. Het partijdige item in het voordeel van beide groepen allochtone leerlingen is het reeds in dit cluster besproken item (1989 Taal nr. 16) dat gaat over het gebruik van de juiste tijd van het werkwoord in een zin gezien de tijdsreferentie in de voorgaande en volgende zinnen. De items die partijdig zijn in het nadeel van deze leerlingen hebben beide betrekking op de vorm van vaste woordcombinaties. Deze items die volgens de IRT-analyses partijdig zijn doen geen afbreuk aan de veronderstelling dat bij kennis van de vorm van woordcombinaties en van conventies op het gebied van de zinsbouw een grotere kans bestaat op bias.

F Referenties (toetsonderdelen Taal en Informatieverwerking)

Van de partijdige items die betrekking hebben op referenties zijn er
- vier in het nadeel van Turkse en Marokkaanse leerlingen en
- drie in het nadeel van Turkse leerlingen.

Er zijn vijf items die bij alle Mantel-Haenszel-analyses – niet significant – in het voordeel zijn van Turkse en/of Marokkaanse leerlingen. Referenties verwijzen naar elementen binnen dezelfde zin of in voorafgaande of volgende zinnen. In de items uit het onderdeel Informatieverwerking wordt veelal gevraagd naar welk woord of naar welke combinatie van woorden een bepaald woord verwijst. In de items die deel uitmaken van de taaltaken wordt veelal gevraagd of een bepaald verwijswoord correct gebruikt wordt. Hierbij moet de leerling het verwijswoord en de referent opsporen en vaststellen of de relatie tussen beide correct is weergegeven. Item 1989 Taal nr. 24 geldt als voorbeeld voor taalitems waarbij de leerling de relatie tussen verwijswoord en referent moet beoordelen.

1989 Taal nr. 24 (taak Taal 2 nr. 4)

1 Hallo Leny, Henk, Ester en Nina,

2 De kinderen van onze school voeren volgende week een musical op voor ouders,
3 andere familieleden en vrienden.
4 Ieder kind uit de drie hoogste groepen doet mee. Ik speel voor reus en mijn vriendin
5 Sanne is de ijskoningin. Ik zou het heel fijn vinden als jullie ook kwamen. Als jullie
6 het leuk vinden, dan moet je ons zo gauw mogelijk bellen.

Wat kun je het beste doen met: *het* (r. 6)?

A Zo laten staan.
B Vervangen door: deze
C Vervangen door: die
D Vervangen door: ze

Ten aanzien van dit item kan verder opgemerkt worden dat de referent van het woord *het* niet direct in de tekst te vinden is, maar uit het grotere geheel van de tekst afgeleid moet worden. Verder is het feit, dat de zin in de tekst ongewijzigd kan blijven en dat de antwoordmogelijkheid *Zo laten staan* gekozen moet worden, ook een mogelijke bron van itembias.
Item 1989 Informatieverwerking nr. 14 is een voorbeeld van een item dat – niet significant – in het voordeel van deze leerlingen is en als contrast kan functioneren voor taalitem 24 uit 1989. Het verwijswoord en de referent staan relatief dicht bij elkaar en er is een directe relatie tussen verwijswoord en referent, bovendien ontbreekt bij dit item de antwoordmogelijkheid *Zo laten staan*.

41 En ik moet zeggen dat ze goed voor de rat zorgen. Het dier krijgt kaas, sla, spek en
42 brood. Als het koud is stoppen ze het in hun binnenzak en nemen het mee naar
43 school waar ze de grotere jongens imponeren.

... stoppen ze het in ... (r. 42)
Wat wordt bedoeld met: *het*?

A de kaas
B de rat
C het brood
D het spek

Dat verwijswoorden bronnen van itembias kunnen zijn sluit aan bij De Jong & Vallen (1989). Zij stellen dat ambigue en moeilijke referenties een bron van bias kunnen zijn. Moeilijke referenties zijn
- verwijzingen waarbij het verwijswoord getalsmatig afwijkt van de referent (bijvoorbeeld: verwijswoord = ze; referent = het gezin),
- verwijzingen waarbij over relatief grote tekstdelen heen wordt verwezen,
- verwijzingen waarbij de schrijver de als bekend veronderstelde informatie niet voor de als nieuw bedoelde informatie plaatst,
- clausale (verwijzing naar – delen uit – een zin) en werkwoordelijke (er wordt verwezen naar een werkwoord meestal in een andere zin) referenties (zie 2.2.2).

Onder de partijdige items komt een referentie voor waarbij het verwijswoord getalsmatig afwijkt van de referent en een referentie waarbij de referent afgeleid moet worden uit de gegeven tekst. Een bijkomende bron van bias kan zijn dat bij de taalitems ook gevraagd wordt of de relatie tussen verwijswoord en referent correct is weergegeven en welk verwijswoord eventueel beter op zijn plaats is. Zowel onder de partijdige items in het nadeel als onder de – niet significante – items in het voordeel van Turkse en Marokkaanse leerlingen komen items voor met verwijzing naar concrete objecten die dicht bij het verwijswoord staan.

Van de items die volgens de IRT-analyses partijdig zijn, behoort er geen item tot het cluster Referenties.

G Spelling (toetsonderdeel Taal)

Van de partijdige items die betrekking hebben op de spelling van werkwoorden en niet-werkwoorden zijn
- twee items in het voordeel van Turkse en Marokkaanse leerlingen,
- acht items in het voordeel van Turkse leerlingen,
- drie items in het voordeel van Marokkaanse leerlingen,
- twee items in het nadeel van Turkse leerlingen.

Het beeld dat partijdige spellingitems in grote meerderheid in het voordeel en niet in het nadeel van Turkse en/of Marokkaanse leerlingen zijn, wordt

versterkt door het feit dat er 19 items zijn die – niet significant – bij alle analyses in het voordeel van Turkse en/of Marokkaanse leerlingen zijn. Bij items die betrekking hebben op spelling moeten leerlingen spellingfouten identificeren in werkwoorden, respectievelijk in woorden met een vast woordbeeld. Van de partijdige items hebben 8 van de 15 items betrekking op de spelling van werkwoorden, dit geldt voor 14 van de 19 items die bij alle analyses – niet significant – in het voordeel van Turkse en/of Marokkaanse leerlingen zijn. Item 1989 Taal nr. 50 is in het nadeel van Turkse leerlingen (De leerlingen moeten bij deze items nagaan of één van de onderstreepte woorden fout gespeld is; als er geen van de onderstreepte woorden fout is, moeten ze D kiezen).

1989 Taal nr. 50 (taak Taal 3 nr. 10)

A Hier volgt een ingelaste uitzending.
B Ontwerpd je moeder die kleren zelf?
C Je kunt de drukte beter vermijden.
D Geen fout.

Bij de beantwoording van item 1989 Taal nr. 50 spelen een aantal taalkundige problemen door elkaar: is 'Ontwerpd' hier een juiste vervoeging van 'ontwerpen', is 'je' een persoonlijk of een bezittelijk voornaamwoord? Achteraf is het niet mogelijk om met zekerheid te zeggen wat de biasbron is. Bovendien moeten we er bij dit soort spellingitems rekening mee houden dat de afleiders A en C invloed kunnen uitoefenen op de keuze voor B.

Item 1989 Taal nr. 53 is in het voordeel van Turkse en Marokkaanse leerlingen.

1989 Taal nr. 53 (taak Taal 3 nr. 13)

A Haggedissen leven op het land.
B We doen het wel met z'n tweeën.
C Uit Italië komen veel olijven.
D Geen fout.

In het totaal zijn er 15 partijdige spellingitems. In 2.2.2 is de verwachting uitgesproken dat de metalinguïstische vaardigheid van tweetalige leerlingen over het algemeen beter is ontwikkeld dan die van monolinguale leerlingen. De genoemde spellingitems doen een beroep op metalinguïstische vaardigheid. Ook moet opgemerkt worden dat de leerbaarheid van spellingregels groot is. Zeker bij de werkwoordspelling gaat het om een transparant en eindig aantal regels die leerlingen moeten leren en moeten leren toepassen. De bijzondere positie die spellingregels voor allochtone leerlingen hebben in vergelijking met andere taalvaardigheden, blijkt ook uit het onderzoek van Kerkhoff (1988). Naar aanleiding van de in haar onderzoek gevonden lage correlaties tussen spelling en andere taalvaardigheden, merkt ze op dat, het correct spellen in een tweede taal beschouwd kan worden als een afzonderlijke vaardigheid, die

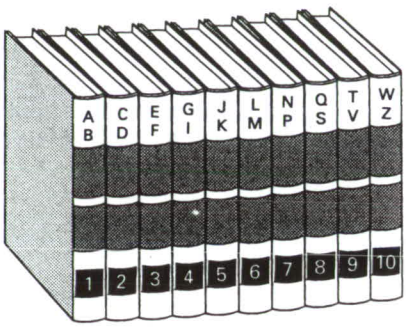
relatief onafhankelijk van andere taalvaardigheden wordt verworven.
De resultaten van de itembiasanalyses lijken te bevestigen dat Turkse en Marokkaanse leerlingen, wellicht door een groter metalinguïstisch bewustzijn, beter dan autochtone leerlingen in staat zijn spellingregels te leren en toe te passen. Bij de spelling van niet-werkwoorden kan verondersteld worden dat allochtone leerlingen minder dan autochtone leerlingen kunnen steunen op een vast woordbeeld en minder ondersteuning zullen ondervinden van de spraakklanken. De resultaten van de uitgevoerde analyses geven echter geen duidelijke grond voor deze bewering. Er zijn relatief veel items over niet-werkwoorden bij alle analyses in het voordeel van Turkse en/of Marokkaanse leerlingen, maar dit voordeel is niet significant ($p<.01$). Het aantal partijdige items over werkwoordspelling is nauwelijks groter dan het aantal items over de spelling van niet-werkwoorden.

Van de items die volgens de IRT-analyses partijdig zijn, behoren er drie tot het cluster Spelling. Alle drie items zijn partijdig in het voordeel van Turkse leerlingen. Twee items hebben betrekking op de spelling van werkwoorden, één op de spelling van niet-werkwoorden. De items die volgens de IRT-analyses partijdig zijn doen geen afbreuk aan de veronderstelling dat spelling een grote kans heeft om biasbron te zijn in het voordeel van Turkse leerlingen.

H De keuze en het gebruik van informatiebronnen (Informatieverwerking)

Van de partijdige items die betrekking hebben op de keuze en het gebruik van informatiebronnen zijn er
- twee in het nadeel van Turkse en Marokkaanse leerlingen,
- twee items zijn in het nadeel van Marokkaanse en
- één item is in het voordeel van Turkse leerlingen.

Bij de partijdige items in het nadeel moeten de leerlingen antwoord geven op de vraag in welk soort informatiebron of bij welke boektitel waarschijnlijk een antwoord te vinden is op een bepaalde vraagstelling. Hierbij moeten de leerlingen de kenmerken van verschillende informatiebronnen kennen (in dit geval: encyclopedie en telefoongids) en ze moeten de kennis over ontsluitingsystemen kunnen toepassen. Bovendien moeten de leerlingen uit de vraagstelling een woord of enkele woorden kiezen die als trefwoorden kunnen functioneren bij de te volgen zoekstrategie. In het item dat partijdig is in het voordeel van Turkse leerlingen komt een Turkse persoon voor: Kamal Atatürk. Voor het juist beantwoorden van dit item is echter geen voorkennis over deze persoon nodig, maar het gaat om de wijze waarop informatie in een encyclopedie gevonden kan worden.



Gert-Jan wil meer weten over de Turkse staatsman Kamal Atatürk.
In welk deel moet hij zoeken?

A in deel 1
B in deel 5
C in deel 8
D in deel 9

Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster De keuze en het gebruik van informatiebronnen.

I Het kennen en kunnen gebruiken van rekenkundige begrippen

Twee partijdige items die betrekking hebben op kennis en toepassing van rekenbegrippen zijn beide in het nadeel van Turkse en Marokkaanse leerlingen. Twee andere items die hierop ook betrekking hebben, zijn – niet significant – in het voordeel van Turkse, respectievelijk Turkse en Marokkaanse leerlingen. Bij de items van dit cluster moeten leerlingen aangeven welke getallen oneven zijn, het begrip deelbaar kunnen gebruiken, het gemiddelde van een aantal getallen kunnen berekenen, de begrippen som, verschil, quotiënt en produkt kunnen hanteren. De items over de begrippen deelbaar en gemiddelde zijn partijdig in het nadeel van Turkse en Marokkaanse leerlingen. Het item over het begrip gemiddelde is het reeds in 7.1.1 afgebeelde item 1987 Rekenen nr. 41. De items over de begrippen oneven, som, verschil, quotiënt en produkt zijn – niet significant – in het voordeel van deze leerlingen. Het gaat in alle gevallen om wat te beschouwen is als schoolse 'rekentaal', die waarschijnlijk aan alle leerlingen onderwezen is.

Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster Het kennen en kunnen gebruiken van begrippen uit het rekenen

J Herleidingen (Rekenen)

Van de twee partijdige items die betrekking hebben op herleidingen is er
– één in het nadeel van Turkse en
– één in het nadeel van Marokkaanse leerlingen.
Tot het cluster Herleidingen behoren ook drie items die – niet significant – in het voordeel zijn van Turkse en/of Marokkaanse leerlingen. Bij Herleidingen gaat het om het transformeren van de ene maateenheid in een andere in een bepaalde context. Item 1989 Rekenen nr. 53, dat partijdig is in het nadeel van Marokkaanse leerlingen, is een voorbeeld hiervan.

1989 Rekenen nr. 53 (taak Rekenen 2 nr. 23)

Iemand moet 3 km wandelen. Zij gaat na hoelang zij doet over honderd meter. Dat duurt ongeveer 1 minuut.
Hoelang zal zij ongeveer onderweg zijn?

A 3 minuten
B 20 minuten
C 30 minuten
D 50 minuten

De leerlingen moeten nagaan hoe vaak 100 meter in 3 kilometer gaat, waarbij kilometers naar meters getransformeerd moeten worden. Verder moeten de leerlingen het aantal keren dat 100 meter in 3 kilometer gaat, vermenigvuldigen met 1 minuut. Uiteraard moet de talige context voldoende aanknopingspunten bieden om te weten welke rekenoperaties verricht moeten worden. Omdat het aantal partijdige items gering is en omdat er ook items over herleidingen – niet significant – in het voordeel van Marokkaans leerlingen zijn, ligt het nog niet voor de hand om te concluderen dat items over herleidingen een bron van itembias zijn.

Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster Herleidingen.

K Omzetten van verhoudingen in procenten en omgekeerd (Rekenen)

Van de drie partijdige items die betrekking hebben op het omzetten van verhoudingen in procenten of omgekeerd is er
– één in het nadeel van Turkse en zijn er
– twee in het nadeel van zowel Turkse als Marokkaanse leerlingen.
Tot dit cluster behoren ook twee items die – niet significant – in het voordeel zijn van Turkse en/of Marokkaanse leerlingen. Tot dit cluster hoort het reeds in 7.1.1 afgebeelde item 1987 Rekenen nr. 27. Item 1989 Rekenen nr. 27 behoort eveneens hiertoe.

1989 Rekenen nr. 27 (taak Rekenen 1 nr. 27)

Op de Arkschool is 1 van elke 2 kinderen lid van een club.
Hoeveel procent is dat?

A $\frac{1}{2}\%$
B $33\frac{1}{3}\%$
C 50%
D 100%

Bij 1989 Rekenen nr. 27 moeten de leerlingen weten dat '1 van elke 2 kinderen' hetzelfde is als 50%. De zinsnede '1 van elke 2 kinderen' is weinig frequent en daarom wellicht biasbron, maar zoals reeds in 7.1.1 is opgemerkt, komt dit taalgebruik in de rekenlessen veelvuldig voor. Opgemerkt wordt dat het bij dit soort formuleringen misschien ook niet altijd duidelijk is wat het referentiepunt voor de percentageberekening is (Wat is 100%?).
In 7.1.1 is reeds aangegeven dat er items zijn die inhoudelijk sterk vergelijkbaar zijn, maar die de ene keer wel en de andere keer niet partijdig zijn. Dit is het geval met twee items uit het onderhavige cluster.
Omdat ook bij dit cluster het aantal partijdige items gering is en omdat er ook items over het transformeren van verhoudingen in procenten – niet significant – in het voordeel van Turkse en/of Marokkaans leerlingen zijn, ligt het nog niet voor de hand om te concluderen dat het omzetten van verhoudingen in procenten op zich een bron van itembias is.

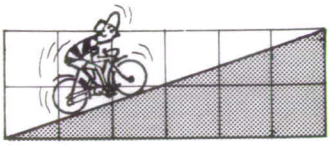
Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster Omzetten van verhoudingen in procenten en omgekeerd.

L Verhoudingen (Rekenen)

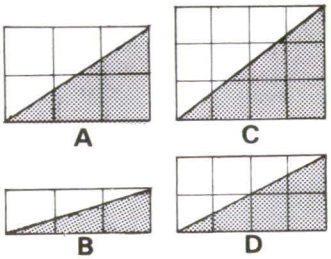
Van de vier partijdige items die betrekking hebben op verhoudingen zijn er
– twee in het nadeel van Turkse en
– twee in het nadeel van zowel Turkse als Marokkaanse leerlingen.
Twee items van dit cluster zijn – niet significant – in het voordeel van Turkse leerlingen. In deze items moeten de leerlingen telkens twee objecten of hoeveelheden aan elkaar relateren en de relatie in de vorm van een bepaalde verhouding beschrijven. Bij de meeste items wordt deze verhouding beschreven met termen als 'het langste', '1 : 2', 'even steil'. Item 1989 Rekenen nr. 28 gaat over het laatste voorbeeld.

1989 Rekenen nr. 28 (taak Rekenen 1 nr. 28)

Voorbeeld



Welke helling hieronder is even steil als de helling uit het voorbeeld?



Het is denkbaar dat de abstractheid en de geringe frequentie van 'is even steil als' de biasbron is. In 7.1.1 is opgemerkt dat items over Verhoudingen een grotere kans maken om een bron van bias te zijn, omdat scholen met relatief veel allochtone leerlingen de leerstof over verhoudingen niet of minder uitgebreid behandelen in vergelijking met andere rekenleerstof. Dit geldt ook voor scholen met meer traditionele rekenmethoden en hieronder zijn de scholen met veel allochtone leerlingen oververtegenwoordigd (Wijnstra, 1988). Bij het rekendomein Verhoudingen is het aannemelijk te veronderstellen dat het summiere of het ontbrekende onderwijsaanbod biasbron is (zie 2.2.4). Gezien het geringe aantal partijdige items in dit cluster moeten we deze veronderstelling met de nodige voorzichtigheid hanteren.

Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster Verhoudingen.

M Rekenitems met relatief veel context en rekenitems met relatief weinig of geen context

Van de zeven partijdige items die behoren tot de categorie rekenitems met *relatief veel context* zijn er
- vier in het nadeel van Turkse leerlingen, is er
- één in het nadeel van Marokkaanse leerlingen en zijn er
- twee in het nadeel van zowel Turkse als Marokkaanse leerlingen.

Bij de items die tot deze categorie behoren, moeten de leerlingen vraagstukjes met meestal veel talige context oplossen waarbij enkele bewerkingen uitgevoerd moeten worden. Item 1987 Rekenen nr. 52 is hiervan een voorbeeld.

1987 Rekenen nr. 52 (taak Rekenen 2 nr. 22)

Per jaar gaf een gezin gemiddeld *f* 1500,– uit aan aardappels en groenten. Ze wilden bezuinigen. Ze huurden een jaar een tuin van 200 m^2 voor *f* 2,– per vierkante meter. De overige onkosten waren *f* 80,–. Ze moesten nog voor een bedrag van *f* 400,– aan aardappelen en groenten in de winkel kopen. De rest kwam uit de tuin.
Hoeveel had dat gezin bespaard met tuinieren?

A *f* 620,–
B *f* 820,–
C *f* 920,–
D *f* 1020,–

De leerlingen moeten uit de tekst opmaken welke bewerkingen ze achtereenvolgens moeten uitvoeren. Door het relatief grote aantal getallen dat in de tekst gegeven is en gebruikt moet worden, doet dit item een groot beroep op taalvaardigheid Nederlands. Opgemerkt wordt dat dit item een groot aantal verwijzingen bevat. Bovendien zijn enkele verwijzingen wellicht complex, omdat het verwijswoord (ze) en de referent (het gezin) getalsmatig afwijken (zie ook het cluster Referenties).

Items met relatief veel context en met laagfrequente woorden die voor het oplossen van het vraagstuk cruciaal zijn, hebben ook een grotere kans om partijdig te zijn. Item 1987 Rekenen nr. 57 is partijdig in het nadeel van zowel Turkse als Marokkaanse leerlingen.

1987 Rekenen nr. 57 (taak Rekenen 2 nr. 27)

Vader koopt een naaimachine. Deze kost *f* 800,– zonder B.T.W. De B.T.W. is 20%.
Hoeveel moet vader betalen inclusief B.T.W.?

A *f* 160,–
B *f* 640,–
C *f* 820,–
D *f* 960,–

Om bij dit item de juiste rekenbewerking te kunnen kiezen (*f* 800,– + *f* 160,– =), moet de betekenis van het woord 'inclusief' voldoende bekend zijn.

Van de zes partijdige items die behoren tot de categorie rekenitems met *relatief weinig of geen context* zijn er
– vier in het voordeel van Marokkaanse leerlingen,
– één item is in het voordeel van Turkse leerlingen en
– één item is in het nadeel van zowel Turkse als Marokkaanse leerlingen.
Tot deze categorie behoren items waarbij leerlingen vraagstukjes met relatief weinig meestal talige context moeten oplossen en waarbij slechts één bewerking

uitgevoerd moet worden. Item 1989 Rekenen nr. 16 is een item met relatief weinig talige context en waarbij één bewerking moet worden uitgevoerd (8000 : 25 =). Dit item is partijdig in het voordeel van Marokkaanse leerlingen.

1989 Rekenen nr. 16 (taak Rekenen 1 nr. 16)

Bert verpakt 8000 bonbons in dozen. In elke doos moeten 25 bonbons.
Hoeveel dozen heeft Bert nodig?

A 32
B 320
C 3200
D 32 000

Bovenstaand item is niet alleen een item met relatief weinig talige context, maar er ontbreken ook laagfrequente woorden die voor het oplossen van het vraagstukje essentieel zijn.

Item 1989 Rekenen nr. 39 is een item zonder talige context, dat partijdig is in het voordeel van Marokkaanse leerlingen.

1989 Rekenen nr. 39 (taak Rekenen 2 nr. 9)

$8{,}5 + \frac{1}{5} =$

A $8\frac{2}{5}$
B 8,6
C 8,7
D 9

Bij de partijdige rekenitems valt een zekere samenhang te constateren tussen de lengte van de talige context van het item en de mate van partijdigheid in het nadeel van Turkse of Marokkaanse leerlingen. Hierbij spelen mogelijk ingewikkelde referenties en laagfrequente woorden, die voor het oplossen van de opgave cruciaal zijn, een grote rol. Bij dergelijke opgaven kan niet volstaan worden met globaal begrip van de tekst, maar is het begrijpen van een enkel woord of de samenhang tussen twee woorden een voorwaarde om het item goed te kunnen beantwoorden. Voor items die partijdig zijn in het voordeel van Turkse en Marokkaanse leerlingen lijkt het omgekeerde te gelden: ze hebben minder talige context, de vraagstelling is minder complex, ze bevatten geen laagfrequente sleutelwoorden en ingewikkelde referenties ontbreken.
Van de items die volgens de IRT-analyses partijdig zijn, behoort er één item (het onderstaande) tot de categorie rekenitem met relatief weinig context: 1989 Rekenen nr. 32. Dit item is niet partijdig volgens de Mantel-Haenszel-analyses.

1989 Rekenen nr. 32 (taak Rekenen 2 nr. 2)

Welke twee getallen liggen even ver van 1 af?

A 0,99 en 1,99
B 1,01 en 1,10
C 0,9 en 1,1
D 0,95 en 1,5

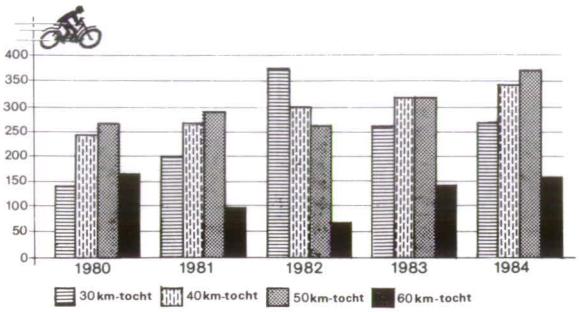
Dit item is in het nadeel van Marokkaanse leerlingen, hetgeen dus afbreuk doet aan de zojuist geuite veronderstelling dat rekenitems met relatief weinig context in het voordeel van Turkse en/of Marokkaanse leerlingen zijn. Het is denkbaar dat de formulering 'ligt even ver af van' voor Turkse en Marokkaanse leerlingen een grotere kans op bias oplevert dan voor autochtone leerlingen.

N Items met een grafische context (Informatieverwerking)

Van de tien partijdige items met relatief veel grafisch materiaal als context zijn er
- zes in het nadeel van Marokkaanse leerlingen,
- drie in het nadeel van Turkse en Marokkaanse leerlingen en
- één item is in het voordeel van Turkse leerlingen.

Bij de items die tot dit cluster behoren, moeten de leerlingen gegevens die gepresenteerd zijn in de vorm van kaarten, tabellen en grafieken identificeren en interpreteren. De grafische presentaties in de partijdige items bestaan uit: de plattegrond van een dierentuin, een thematische geografische kaart, kruistabellen, een staaf- en een cirkelgrafiek. Opgemerkt moet worden dat de meeste items uit dit cluster ook veel talige informatie bevatten die voor het oplossen van het item van groot belang is. Item 1989 Informatieverwerking 40 is een voorbeeld van een partijdig item in het nadeel van zowel Turkse als Marokkaanse leerlingen.

Aantal deelnemers aan de fietstochten in Lutterveen



Voor welke afstand geldt dat het aantal deelnemers elk jaar is toegenomen?

A voor de 30 km-tocht
B voor de 40 km-tocht
C voor de 50 km-tocht
D voor de 60 km-tocht

De leerlingen moeten de lengte van de verschillende staven vergelijken en op basis hiervan een bepaalde trend (toename) onderkennen. Vervolgens moeten ze nagaan welke fietstocht met een bepaalde arcering bedoeld wordt. Bovendien moeten de leerlingen in de begrippen 'afstand','geldt' en 'toegenomen' voldoende aanknopingspunten vinden om het item goed te beantwoorden. Zowel de grafische als de talige elementen in het item kunnen de bron van itembias vormen.

Van de items die volgens de IRT-analyses partijdig zijn, behoort geen enkel item tot het cluster Items met grafische context.

7.1.3 Overeenstemming tussen de inhoudsanalyse van items die volgens de Mantel-Haenszel- en de IRT-procedure partijdig zijn

Uit 7.1.2 volgt dat bij een aantal clusters de resultaten van de inhoudelijke analyse van items die partijdig zijn volgens zowel de Mantel-Haenszel- als de IRT-procedure overeenstemmen. De resultaten van de inhoudsanalyses zijn bij de IRT-procedure op minder items gebaseerd dan bij de Mantel-Haenszel-procedure. Bij een groot aantal clusters zijn er alleen items volgens de Mantel-Haenszel-procedure partijdig. Bij deze clusters kunnen minder stellig uitspraken gedaan worden over bronnen van itembias, zeker wanneer het aantal items waarop de conclusies gebaseerd zijn, gering is.
Wanneer echter beide itembiasdetectieprocedures in dezelfde richting wijzen

inzake bronnen van itembias, dan is de mate van zekerheid hierover groter. Deze overeenstemming geldt met name ten aanzien van de volgende bronnen van itembias:

- **B** Tekstbegrip (begrijpend lezen):
  Items die naar globaal tekstbegrip (macroniveau) vragen, hebben kans op itembias in het voordeel van Turkse en Marokkaanse leerlingen. Items die vragen naar de betekenis van een specifiek woord of een specifieke zin, door een woordelijke of geparafraseerde herhaling van expliciet in de tekst gegeven informatie te vragen (tekstbegrip op meso-/microniveau), hebben kans op itembias in het nadeel van deze leerlingen.
- **C** Woordkennis en kennis van woordcombinaties
  Items die vragen naar de betekenis van moeilijke woorden, waarvan de betekenis niet of moeilijk uit de context kan worden afgeleid, hebben een kans op itembias in het nadeel van Turkse en Marokkaanse leerlingen.
- **E** Correct taalgebruik
  Items die betrekking hebben op de kennis van de vorm van vaste woordcombinaties en conventies op het gebied van de zinsbouw, hebben een kans partijdig te zijn in het nadeel van Turkse en Marokkaanse leerlingen .
- **G** Spelling
  Items die vragen spelfouten in werkwoorden en in woorden met een vast woordbeeld aan te geven, hebben een kans op itembias in het voordeel van deze leerlingen.

De mate van zekerheid inzake bronnen van itembias is bij een deel van de clusters geringer, omdat de items alleen partijdig zijn volgens de Mantel-Haenszel-procedure. Dit geldt ook voor de items uit het cluster Referenties. Toch kan er hier sprake zijn een grotere mate van zekerheid inzake bronnen van itembias, omdat referenties bij twee clusters genoemd worden als een bron van itembias: Referenties en Rekenitems met relatief veel context.

De bovenstaande mogelijke bronnen van itembias zijn te beschouwen als de eerste resultaten van de inhoudsanalyse van partijdige items. Dit resultaat is gebaseerd op de gemeenschappelijkheden in de analyses van de afzonderlijke projectmedewerkers. Om meer aanwijzingen over bronnen van itembias te verkrijgen zijn ook niet bij het project betrokken experts gevraagd partijdige items op mogelijke biasbronnen te beoordelen (7.2).

7.2 Oordelen van experts over mogelijke bronnen van itembias

De in 7.2.1 beschreven conclusies moeten gezien de beperkingen, die in 7.1.1 genoemd zijn, als voorlopig beschouwd worden. In de volgende fase van het onderzoek zijn niet bij het project betrokken experts gevraagd een aantal partijdige items in het voordeel, respectievelijk nadeel van Turkse en Marokkaanse leerlingen inhoudelijk te beoordelen en aan te geven welke itemelementen voor deze leerlingen naar hun oordeel moeilijk/makkelijk zijn. De oordelen van deze deskundigen afkomstig uit diverse wetenschapsdisciplines en de onderwijspraktijk zijn enerzijds onderling vergeleken en anderzijds vergeleken met de oordelen van de projectmedewerkers. In 7.2.1 wordt de opzet van dit onderzoek naar expert-oordelen beschreven, terwijl in 7.2.2 de

resultaten ervan vermeld staan. Van de Waal doet in haar door Vallen en Uiterwijk begeleide doctoraalscriptie uitgebreid verslag van het onderzoek naar de expert-oordelen (Van de Waal, 1992). In dit hoofdstuk worden de hoofdzaken daaruit aangehaald en wordt verslag gedaan van een nadere analyse van de gegevens.

7.2.1 Opzet van het onderzoek naar de oordelen van experts

Het doel van dit deelonderzoek luidt:
- bepalen in welke mate de oordelen van de experts aansluiten bij de oordelen van de projectmedewerkers inzake de bronnen van itembias in partijdige items;
- bepalen in welke mate er overeenkomst bestaat tussen de oordelen van de experts over de bronnen van bias in partijdige items.

De experts zijn items voorgelegd, die bij drie Mantel-Haenszel-analyses partijdig zijn in het voordeel, respectievelijk het nadeel van Turkse en/of Marokkaanse leerlingen (zie 6.1.3). Om het aantal items in het voordeel van beide groepen leerlingen te verhogen bestaat een deel van de in totaal 84 bevraagde items uit items die bij alle Mantel-Haenszel-analyses – niet significant – in het voordeel van deze leerlingen zijn. De totale itemlijst bestaat uit 37 taalitems, 31 rekenitems en 16 informatieverwerkingitems.
De experts is gevraagd te beoordelen welke items moeilijker dan wel makkelijker voor allochtone leerlingen zijn en welke oorzaken daarvoor te geven zijn. De experts werd bovendien gevraagd ten aanzien van allochtone leerlingen onderscheid te maken tussen Turkse en Marokkaanse leerlingen.
De experts wisten vooraf niet of een item partijdig is in het voor- of nadeel van deze leerlingen.
De experts is niet gevraagd om onderscheid te maken tussen partijdige items (in het voor- of nadeel van deze leerlingen) en moeilijke/makkelijke items, omdat de experts dan ook duidelijk voor ogen zouden moeten hebben ten aanzien van welke vaardigheid allochtone en autochtone leerlingen een gelijk prestatieniveau hebben (zie 1.2.2). Om dit probleem te vermijden is gekozen voor de dimensie moeilijk-makkelijk. Deze dimensie geeft op een andere wijze aan (niet gecorrigeerd voor de 'gezuiverde totaalscore') of een item in het voordeel of in het nadeel van allochtone leerlingen is en is in feite een variatie op de dimensie partijdig in het voor- of nadeel.

De experts zijn vertegenwoordigers uit de onderwijspraktijk en uit verschillende wetenschappelijke disciplines. De experts uit de laatste groep hebben allen ervaring in en/of gebleken belangstelling voor toets- en/of curriculumontwikkeling, al dan niet voor allochtone leerlingen. Omdat de mogelijkheid bestaat dat allochtone experts op linguïstisch en cultureel terrein andere informatie verschaffen dan autochtone, zijn ook allochtone experts in het onderzoek betrokken. Tabel 7.2 geeft de verdeling over de onderscheiden categorieën weer.

Tabel 7.2 Verdeling van de experts over de onderscheiden categorieën

Experts	Allochtoon	Autochtoon	Totaal
onderwijsgevenden	2	1	3
taalkundigen	2	7	9
onderwijskundigen	1	1	2
toetsconstructeurs (Cito)	2	2	
totaal	5	11	16

7.2.2 Resultaten van het onderzoek naar de oordelen van experts

De 16 experts hebben nadat ze de 84 items hadden bestudeerd, per item aangegeven of een item naar hun oordeel voor allochtone leerlingen moeilijker of makkelijker is dan voor autochtone leerlingen. De experts is ook verzocht zich uit te spreken over vraag waarom een item of een reeks van items moeilijker dan wel makkelijker zou zijn voor allochtone leerlingen. In Van de Waal (1992) is een samenvatting van de gesprekken opgenomen.
Uit een nadere analyse van de gegevens in Van de Waal (1992) blijkt dat de experts er niet in alle opzichten in geslaagd zijn aan te geven of een item moeilijker dan wel makkelijker is voor allochtone leerlingen. Als het om moeilijke, respectievelijk makkelijke items gaat, maken vrijwel alle experts bijvoorbeeld geen onderscheid tussen de volgende twee items, terwijl de eerder beschreven itembias-analyses daar wel aanleiding toe geven.

1989 Rekenen nr. 27 (taak Rekenen 1 nr. 27)

Op de Arkschool is 1 van elke 2 kinderen lid van een club.
Hoeveel procent is dat?

A $\frac{1}{2}\%$
B $33\frac{1}{3}\%$
C 50%
D 100%

Item 1989 Rekenen nr. 27 is partijdig in het nadeel van zowel Turkse als Marokkaanse leerlingen.
Het volgende item, 1989 Rekenen nr. 57, is – niet significant – in het voordeel van beide groepen leerlingen.

1989 Rekenen nr. 57 (taak Rekenen 2 nr. 27)

De olieprijs daalde van 20 tot 15 dollar per vat.
Hoeveel procent daalde de prijs?

A 3%
B 4%
C 5%
D 50%

Vrijwel alle experts beschouwen beide items als gelijkwaardig inzake de dimensie moeilijk-makkelijk. Eén expert merkt van item 1989 Rekenen nr. 27 op dat de referentie 'dat' en de formulering '1 van elke 2' moeilijke aspecten voor allochtone leerlingen zijn. Twee experts merken van item 1989 Rekenen nr. 57 op dat bij dit item voor allochtone leerlingen de moeilijkheid in de complexiteit van de opdracht zit.
Er is nagegaan in welke mate de experts er in slagen aan te geven of een item moeilijker is voor allochtone dan voor autochtone leerlingen. Per item is vastgesteld of meer dan de helft van de respondenten zegt dat het item moeilijker is voor allochtone leerlingen en of het item inderdaad partijdig is in het nadeel van allochtone leerlingen. Gebleken is dat er op dit punt bij 31 van de 84 items (37%) geen overeenstemming bestaat tussen het oordeel van de meerderheid van de experts en de richting (voor- of nadeel) van de partijdigheid.
Tevens is vastgesteld hoe hoog de samenhang is tussen het aantal experts dat zegt dat een item moeilijker is voor allochtone leerlingen en de mate van itembias (aantal keren in het voordeel, respectievelijk nadeel van allochtone leerlingen). De correlatie tussen het aantal experts dat zegt dat het item moeilijker is en de mate van itembias is niet hoog te noemen: $r = .30$ ($p<.01$).
Feit is natuurlijk dat de dimensies partijdigheid in het voor- of nadeel en makkelijk-moeilijk niet hetzelfde zijn, maar het is niet aannemelijk te veronderstellen dat dit onderscheid de geringe trefzekerheid van de experts verklaart.
Verder is gebleken dat de experts niet of nauwelijks onderscheid maken tussen moeilijke/makkelijke items voor Turkse en Marokkaanse leerlingen.

De oordelen van de experts over de dimensie moeilijk-makkelijk blijken onderling sterk overeen te stemmen en sluiten in hoge mate aan bij de bevindingen van de projectmedewerkers (zie 7.1.2). De experts zijn echter minder dan de projectmedewerkers van mening dat items die betrekking hebben op het herstellen van fouten in de structuur van een tekst (weghalen van redundanties in een tekst, het verwijderen van fouten in de opbouw van een tekst) problematisch kunnen zijn voor allochtone leerlingen. De experts benadrukken ook dat items over de woordkennis en de kennis van idiomatische uitdrukkingen een grote kans maken moeilijk te zijn voor allochtone leerlingen. Daarnaast geven de experts aan dat items die een beroep doen op cultureel bepaalde voorkennis waarover allochtone leerlingen niet beschikken, moeilijk kunnen zijn voor deze leerlingen. Deze voorkennis kan bijvoorbeeld een rol

spelen bij de items waarbij een uitvoerige tekst als contextmateriaal fungeert. Bij een groot aantal items zijn de experts van oordeel dat de opgaven geschreven zijn vanuit een Nederlandse cultuurkennis en daardoor minder toegankelijk zijn voor veel allochtone leerlingen.

Opmerkelijk is dat experts niet erg trefzeker zijn in het onderscheiden van items die in het voor- dan wel in het nadeel van Turkse en Marokkaanse leerlingen zijn, maar toch in hoge mate aansluiten bij de oordelen van de project-medewerkers over bronnen van itembias. Het aantal voorgelegde items wel of niet significant in het voordeel van deze leerlingen is over het algemeen geringer dan het aantal items in het nadeel.
Partijdige items in het voordeel van Turkse en/of Marokkaanse leerlingen hebben vaak betrekking op dezelfde bron van itembias als de items die partijdig zijn in het nadeel, maar zijn aanzienlijk geringer in aantal (zie 7.1.2). Ze zwakken in feite de conclusies af die over bronnen van itembias getrokken kunnen worden. Itemclusters die alleen partijdige items in het nadeel van Turkse en/of Marokkaanse leerlingen kennen, zijn (zie 7.1.3): Woordkennis en kennis van woordcombinaties en Rekenitems met relatief veel context. Bij het cluster Spelling is de situatie omgekeerd: de meeste partijdige spellingitems zijn in het voordeel van Turkse en/of Marokkaanse leerlingen en slechts enkele items zijn in het nadeel van deze leerlingen.

7.2.3 Conclusies uit het onderzoek naar de oordelen van experts

De experts blijken niet erg trefzeker te zijn in het opsporen van partijdige items in het voor-, respectievelijk nadeel van Turkse en Marokkaanse leerlingen. Natuurlijk zijn de dimensies partijdigheid in het voor- of nadeel en makkelijk-moeilijk niet identiek, maar het is niet aannemelijk te veronderstellen dat dit verschil de geringe trefzekerheid verklaart.
De experts maken geen onderscheid tussen moeilijke of makkelijke items voor Turkse dan wel Marokkaanse leerlingen.
De oordelen van de experts over bronnen van moeilijkheid corresponderen onderling in hoge mate en sluiten over het algemeen aan bij de bevindingen van de projectmedewerkers (zie 7.1.2). Daarnaast benadrukken de experts dat verschillen in cultureel bepaalde voorkennis tussen de autochtone en allochtone leerlingen zeer waarschijnlijk een probleem is. Deze voorkennis kan bijvoorbeeld een rol spelen bij de items waarvoor contextmateriaal wordt gebruikt. Zo kunnen allochtone leerlingen minder vertrouwd zijn met het onderwerp dat in een tekst aan de orde wordt gesteld. De experts zijn van oordeel dat veel opgaven geschreven zijn vanuit de Nederlandse cultuurkennis en daardoor voor veel allochtone leerlingen minder herkenbaar zijn.

7.3 Een hardop-denken-experiment voor het opsporen van mogelijke bronnen van itembias

In aanvulling op de tot dusver beschreven pogingen om bronnen van itembias op te sporen hebben de projectmedewerkers ook allochtone en autochtone leerlingen uit groep acht van het basisonderwijs bij het project betrokken. Met een kleinschalig hardop-denken-experiment is nagegaan hoe vaak allochtone en

autochtone leerlingen bij partijdige items een fout antwoord geven ten gevolge van een element in het item dat voorlopig als bron voor itembias is aangewezen. Tevens is onderzocht hoe vaak bij gemanipuleerde items door de item-manipulatie een goed antwoord wordt gegeven. Gemanipuleerde items zijn items waarbij het itemelement dat als potentiële biasbron is aangewezen (bijvoorbeeld: 'Hoeveel moet hij betalen *inclusief* B.T.W.'), is vervangen door een itemelement waarvan verwacht wordt dat het geen bias veroorzaakt (bijvoorbeeld: 'Hoeveel moet hij betalen *met* B.T.W.').
De allochtone en autochtone leerlingen moesten eerst de hen voorgelegde oorspronkelijke, respectievelijk gemanipuleerde items goed bestuderen, daarna moesten ze het naar hun oordeel goede antwoord aankruisen. Tot slot moesten ze zo uitgebreid en nauwkeurig mogelijk aangeven hoe ze aan hun antwoord gekomen waren. In de redeneringen van de leerlingen is getracht aanwijzingen te vinden omtrent bronnen van itembias (zie Coenen & Vallen, 1991).

Er mag in verband met de constructvaliditeit verondersteld worden dat in bijvoorbeeld rekenitems de benodigde talige context communaal is en niet zo moeilijk is dat hij discrimineert tussen allochtone en autochtone leerlingen. Indien de talige context bij rekenitems wel zou discrimineren tussen allochtone en autochtone leerlingen, dan moet dat itemelement vervangen worden door een talig element dat wel communaal is. Als partijdige items op de juiste wijze gemanipuleerd worden, is te verwachten dat allochtone leerlingen minder fouten in de items maken en dat allochtone en autochtone leerlingen met hetzelfde vaardigheidsniveau ongeveer even veel opgaven goed maken. Wellicht dat de items dan bij beide groepen leerlingen beter de vaardigheid meten die ze beogen te meten. Van belang hierbij is wel dat het oorspronkelijke en het gemanipuleerde item dezelfde vaardigheid meten.

7.3.1 Opzet van het hardop-denken-experiment

Ten behoeve van dit experiment zijn de items uit de toetsonderdelen Taal, Rekenen en Informatieverwerking van de Eindtoets Basisonderwijs 1987 genomen die bij de Mantel-Haenszel-analyses sterk partijdig zijn in het nadeel van Turkse en/of Marokkaanse leerlingen.
Uit de partijdige items zijn alleen de items gekozen die zo veranderd kunnen worden, dat het gemanipuleerde item nog steeds hetzelfde pretendeert te meten als het oorspronkelijke item. Dit heeft tot gevolg dat de items slechts minimaal gemanipuleerd werden. Bij een gemanipuleerd item werd het itemelement dat door de projectmedewerkers voorlopig als biasbron was aangewezen, vervangen door een itemelement waarvan verwacht werd dat het geen bias veroorzaakte. In 7.3.2 worden bij de bespreking van de resultaten van het hardop-denken-experiment, drie voorbeelden gegeven van oorspronkelijke en gemanipuleerde items.
De talige manipulatie had vooral betrekking op onnodig moeilijke woorden, complexe grammaticale en/of ambigue constructies en op impliciete zin- en tekststructuren. Bovendien hebben manipulaties plaatsgevonden in grafische contexten om veronderstelde onduidelijkheden in tekeningen, kaarten en tabellen te verwijderen.
Uit de Eindtoets Basisonderwijs 1987 zijn acht reken-, vier informatie-verwerking- en vijf taalitems geselecteerd. De gekozen items kunnen niet

beschouwd worden als een representatieve steekproef uit alle items van de Eindtoets Basisonderwijs, omdat alleen items zijn geselecteerd, die na de itemmanipulatie nog steeds hetzelfde beogen te meten als het oorspronkelijke item. De 17 items zijn in toetsversie A in de oorspronkelijke vorm getoetst en in toetsversie B in de gemanipuleerde vorm. Deze items zijn voorgelegd aan 22 allochtone en 22 autochtone leerlingen uit groep acht van vijf basisscholen. Tabel 7.3 geeft de verdeling van de leerlingen over de beide toetsversies weer.

Tabel 7.3 Verdeling van de leerlingen over de beide toetsversies

Toetsversie	Allochtonen	Autochtonen
Toetsversie A: 17 oorspronkelijke items	11	11
Toetsversie B: 17 gemanipuleerde items	11	11

In overleg met de leerkrachten is telkens bij elke allochtone leerling een autochtone leerling geselecteerd die vergelijkbaar was op factoren die van belang zijn voor schoolsucces, zoals sociaal-economische achtergrond, taalvaardigheid Nederlands, rekenvaardigheid, motivatie, doubleergeschiedenis, Eindtoetsscore en schoolkeuze-advies van de basisschool. Alle allochtone leerlingen moesten voldoen aan twee criteria. Ze moesten thuis een allochtone taal spreken en ze moesten minimaal vanaf groep drie het Nederlandse basisonderwijs volgen.
Bij de verdeling van de leerlingen over de beide toetsversies is er op gelet dat het prestatieniveau van de beide groepen leerlingen zoveel mogelijk vergelijkbaar is. Onder de leerlingen die toetsversie A dan wel B maken, zitten ongeveer evenveel leerlingen met een LBO- of MAVO- of HAVO-advies. Zowel bij de allochtone als bij de autochtone leerlingen is het aantal jongens en meisjes gelijk. De groep allochtone leerlingen bestaat uit elf Turkse, acht Marokkaanse leerlingen, één Chinese, één Antilliaanse en één Braziliaanse leerling. De Turkse en Marokkaanse leerlingen zijn nagenoeg gelijk verdeeld over de beide toetsversies.

De leerlingen kregen, nadat de items waren voorgelegd, de opdracht om elk item goed te bestuderen en het goede antwoord te kiezen. Daarna moesten ze zo uitgebreid en precies mogelijk vertellen hoe ze de taakstelling van elk item opgelost hadden. Als de leerling het item fout oploste of een onduidelijke toelichting gaf, werd er door de projectmedewerker verder gevraagd om er achter te komen of de leerling de bedoelde vaardigheid beheerste. De gesprekken werden op audio-cassette opgenomen. Bij het afluisteren van de opnames werd er vooral op gelet of de itemelementen, die voorlopig als biasbron waren aangewezen, voor de leerlingen inderdaad een probleem vormde bij het oplossen van het item.

7.3.2 Resultaten van het hardop-denken-experiment

De gemiddelde scores van de allochtone en autochtone leerlingen die toetsversie A, respectievelijk toetsversie B maakten, geven een eerste indicatie voor het effect dat de itemmanipulatie heeft gehad. Het betreft slechts een indicatie, omdat het aantal leerlingen (n=11) en het aantal items (k=17) gering is en omdat niet met volledige zekerheid gezegd kan worden of de onderscheiden subgroepen even vaardig zijn in hetgeen de items beogen te meten. In tabel 7.4 staat per onderscheiden subgroep het gemiddeld percentage goed gemaakte items.

Tabel 7.4 Gemiddeld percentage goed gemaakte items per subgroep

Toetsversie	Allochtonen (n=11)	Autochtonen (n=11)
Taal		
Toetsversie A: k=5	58.2	78.2
Toetsversie B: k=5	67.3	87.3
Verschil B–A	9.1	9.1
Rekenen		
Toetsversie A: k=8	36.4	55.7
Toetsversie B: k=8	53.4	58.0
Verschil B–A	17.0	2.3
Informatieverwerking		
Toetsversie A: k=4	52.3	84.1
Toetsversie B: k=4	72.7	84.1
Verschil B–A	20.4	0
Totaal		
Toetsversie A: k=17	46.5	69.0
Toetsversie B: k=17	62.0	72.7
Verschil B–A	15.5	3.7

Toelichting:
Toetsversie A = de toetsversie met de oorspronkelijke items
Toetsversie B = de toetsversie met de gemanipuleerde items

De allochtone leerlingen die de gemanipuleerde items gemaakt hebben, maken in totaal gemiddeld 15.5% meer items goed dan de allochtone leerlingen die de oorspronkelijke items maakten. Bij de autochtone leerlingen is het verschil slechts 3.7%. Het verschil tussen het gemiddeld percentage goed van de allochtone en autochtone leerlingen die de oorspronkelijke toetsversie maakten, is 22.5%. Het verschil tussen het gemiddeld percentage goed van de allochtone en autochtone leerlingen die de gemanipuleerde toetsversie maakten, is 10.7%. Het verschil tussen de gemiddelde score van de allochtone en autochtone leerlingen die de oorspronkelijke toetsversie maakten, is niet significant voor de leerlingen met een een HAVO-advies, maar wel significant voor de leerlingen

met een MAVO- ($p<.05$) en LBO-advies ($p<.01$). Het verschil tussen beide groepen leerlingen op de gemanipuleerde versie is alleen significant voor de leerlingen met een LBO-advies ($p<.05$).
De gegevens in tabel 7.3 geven aan dat de allochtone leerlingen meer geprofiteerd hebben van de itemmanipulaties dan de autochtone leerlingen. De itemmanipulaties lijken differentiële effecten te hebben voor de items uit de drie toetsonderdelen van de Eindtoets Basisonderwijs. Bij het onderdeel Informatieverwerking maken de allochtone leerlingen alle gemanipuleerde items beter dan de oorspronkelijke items. Ook drie van de acht rekenitems zijn door allochtone leerlingen in de gemanipuleerde versie beter gemaakt dan in de oorspronkelijke versie. Bij de overige rekenitems zijn er nauwelijks verschillen. Bij drie van de vijf taalitems worden eveneens de gemanipuleerde items door de allochtone leerlingen beter gemaakt, maar bij de andere twee items worden de oorspronkelijke items beter gemaakt.
Wanneer we kijken naar de gemiddelde scoreverschillen tussen de oorspronkelijke en de gemanipuleerde taalitems dan lijken de allochtone leerlingen in gelijke mate als de autochtone leerlingen te profiteren van de itemmanipulaties. Bij de items over rekenen en informatieverwerking lijken de allochtone leerlingen meer voordeel te hebben van de itemmanipulaties dan de autochtone leerlingen. De gemiddelde scores per subgroep geven aan dat de gemanipuleerde taalitems gemakkelijker zijn geworden voor iedereen en dat bij de reken- en informatieverwerkingitems de bronnen van itembias voor allochtone leerlingen inderdaad grotendeels verwijderd zijn.
Het is denkbaar dat de gemanipuleerde items voor allochtone leerlingen meer constructvalide zijn dan de oorspronkelijke items. Het is evenwel niet zeker of de gemanipuleerde items op dezelfde vaardigheid een beroep doen als de oorspronkelijke (partijdige) items beogen te doen. Omdat niet nagegaan kan worden welke items wel of niet een eendimensionele schaal vormen, is niet uitgesloten dat de gemanipuleerde items in feite een andere dan de beoogde vaardigheid meten.

Met de inhoudelijke protocol-analyses zijn vooral items opgespoord die door de allochtone leerlingen vaker fout gemaakt werden dan door autochtone leerlingen, maar waarbij de protocollen aanleiding geven te veronderstellen dat de leerling de te meten vaardigheid wel beheerste. Tijdens de gesprekken kon niet altijd goed vastgesteld worden of de leerlingen de vaardigheid beheersten, die het item beoogt te meten. Bij items die in afzonderlijke concrete stappen opgelost moeten worden, konden deze (deel)vaardigheden apart bevraagd worden, maar dat is minder eenvoudig bij items waarbij dat niet het geval is. Bij de nu volgende bespreking van de resultaten wordt als voorbeeld uit elk toetsonderdeel eerst één item uitgebreid besproken. Van de overige items uit het betreffende toetsonderdeel wordt een samenvatting van de resultaten gegeven.

Item 1987 Taal nr. 40 wordt eerst in de oorspronkelijke versie gepresenteerd, daarna in de gemanipuleerde versie.

Oorspronkelijke versie

31 Bijna gelijktijdig schieten Kokkie en Appel de ruimte in. Ze komen aan de voet van
32 een zandberg neer. De bemanning is er zonder kleerscheuren afgelopen. Alleen
33 Appel heeft een paar blauwe plekken opgelopen.

Wat kun je het beste doen met: *afgelopen.* (r. 32)?

A Zo laten staan.
B Vervangen door: afgekomen.
C Vervangen door: afgeraakt.
D Vervangen door: afgevallen.

Bij de gemanipuleerde versie wordt niet gevraagd een passage uit een tekst te verbeteren, maar wordt de leerling gevraagd om in een zin het ontbrekende woord in te vullen.

Gemanipuleerde versie

De autobestuurder was er bij het ongeluk gelukkig zonder kleerscheuren ....

Welk woord past hier het beste?

A afgekomen
B afgelopen
C afgeraakt
D afgevallen

De oorspronkelijke versie wordt door drie allochtone en twee autochtone leerlingen fout gemaakt. Bij de gemanipuleerde versie wordt door één allochtone en door twee autochtone leerlingen een fout antwoord gegeven. Bij de andere taalitems die eveneens woordkennis beogen te meten, wordt bij twee items in de gemanipuleerde versie wellicht niet meer dezelfde vaardigheid gemeten als in de oorspronkelijke versie. Bij één item is de uitdrukking 'wonderlijke genoeg' vervangen door 'vreemd genoeg' en bij de andere items is de context zo aangepast, dat de gevraagde betekenis van het woord 'bijtijds' veel meer dan in de oorspronkelijke versie afgeleid kan worden uit de context. De twee overige taalitems zijn in de gemanipuleerde versie om onduidelijke redenen moeilijker dan in de oorspronkelijke versie.

Item 1987 Rekenen nr. 57 wordt eerst in de oorspronkelijke versie gepresenteerd, daarna in de gemanipuleerde versie. Het gemanipuleerde deel van het item is vet gedrukt.

Oorspronkelijke versie

Vader koopt een naaimachine. Deze kost *f* 800,– zonder B.T.W. De B.T.W. is 20%. Hoeveel moet vader betalen inclusief B.T.W.?

A *f* 160,–
B *f* 640,–
C *f* 820,–
D *f* 960,–

Gemanipuleerde versie

Vader koopt een naaimachine. Deze kost *f* 800,– zonder B.T.W. De B.T.W. is 20%.
Wat moet vader voor de naaimachine betalen met B.T.W?

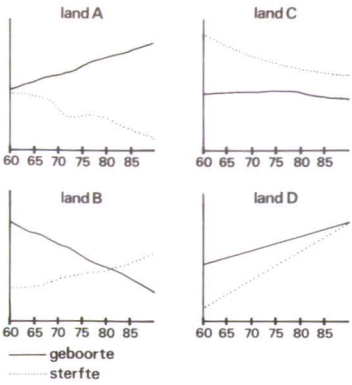
A *f* 160,–
B *f* 640,–
C *f* 820,–
D *f* 960,–

De oorspronkelijke versie wordt door zes van de elf allochtone leerlingen fout beantwoord en door twee van de elf autochtone. Bij de gemanipuleerde versie geven twee allochtone en één autochtone leerling een fout antwoord. Uit de protocol-analyse blijkt dat bij twee allochtone en bij één autochtone leerling het woord 'inclusief' de oorzaak van de fout is. Een andere allochtone leerling heeft ook problemen met 'inclusief', maar nog meer met de vaardigheid in het berekenen van procenten. Als in plaats van 'inclusief' de aanduiding 'met' zou zijn gebruikt, maakt dat (naar het eigen oordeel van de leerlingen) voor twee allochtone leerlingen geen verschil en zou dat voor vier andere allochtone en voor twee autochtone leerlingen makkelijker zijn geweest.
Bij de overige rekenitems levert eenmaal een tekening problemen op. Ook talige aspecten van de rekenitems lijken problemen te veroorzaken. In de vraag 'Welk bootje heeft in verhouding tot zijn lengte de langste mast?' wordt 'zijn lengte' vaak opgevat als de lengte van de mast. De aanduiding 'overige onkosten' blijkt niet altijd bij iedereen bekend te zijn en de aanduiding 'half procent' wordt begripsmatig verward met 'de helft'.

Item 1987 Informatieverwerking nr. 40 wordt eerst in de oorspronkelijke versie gepresenteerd, daarna in de gemanipuleerde versie. Het gemanipuleerde deel van het item is vet gedrukt.

Oorspronkelijke versie

Sterfte- en geboortecijfers in vier landen



Van welk land kan men zeggen dat het aantal geboorten toeneemt en het aantal sterfgevallen afneemt?

A van land A
B van land B
C van land C
D van land D

Gemanipuleerde versie

Zie de grafieken 'Sterfte- en geboortecijfers in vier landen' in de oorspronkelijke versie van het item

In welk land is tussen 1960 en 1985 het aantal geboorten gestegen en het aantal sterfgevallen gedaald?

A van land A
B van land B
C van land C
D van land D

De oorspronkelijke versie wordt door vijf van de elf allochtone leerlingen fout beantwoord en door geen enkele autochtone leerling. Bij de gemanipuleerde versie geven twee allochtone en één autochtone leerling een fout antwoord. Uit de protocol-analyse blijkt dat twee allochtone leerlingen de legenda niet meteen opmerken, omdat die erg dicht op de grafieken staat. Vier allochtone leerlingen hebben problemen met de woorden 'toenemen' en 'afnemen'. Eén van deze leerlingen, die denkt dat afnemen 'iemand zijn spullen afpakken' betekent en toenemen 'pakken', komt via de redenering dat de andere grafieken 'raar' zijn

toch op het goede antwoord uit. De betekenis van 'toenemen' en 'afnemen' is voor één allochtone leerling waarschijnlijk de belangrijkste foutenbron en voor een andere wellicht mede oorzaak van de fout. De vier leerlingen die moeite hebben met de begrippen 'afnemen' en 'toenemen', begrijpen het begrippenpaar 'stijgen' en 'dalen' wel. Bij de overige informatieverwerkingitems spelen vooral onduidelijke grafische presentaties een rol (zie Coenen & Vallen, 1991).

7.3.3 Conclusies uit het hardop-denken-experiment

Voordat we de resultaten van het hardop-denken-experiment overzien, moet herhaald worden dat de opzet van het experiment zijn beperkingen kent:
- het aantal leerlingen (n=11) en het aantal items (k=17) is gering;
- het is niet zeker of de gemanipuleerde items op dezelfde vaardigheid een beroep doen als de oorspronkelijke (partijdige) items beogen te doen;
- tijdens de gesprekken kon niet altijd goed vastgesteld worden of de leerlingen de vaardigheid die het item beoogt te meten, beheersten;
- er kan niet met zekerheid gesteld worden of de allochtone en autochtone leerlingen, die de oorspronkelijke, respectievelijk de gemanipuleerde items maakten, even vaardig zijn in hetgeen de items beogen te meten.

Met inachtneming van bovenstaande beperkingen kan gezegd worden, dat het hardop-denken-experiment aanwijzingen geeft dat biasbronnen voor een groot deel op het gebied van woordgebruik en impliciete zins- en tekstverbanden gezocht moeten worden. Daarnaast leiden ongebruikelijke uitdrukkingen ('niet te laat') en woordvormgelijkenis ('half procent'/'helft') tot problemen. Verder zijn grafische onduidelijkheden verwarrend. De complexiteit van de items speelt eveneens een rol. Complexe items vereisen doorgaans meer context en voor het oplossen van het item moet de leerling meestal een aantal tussenstappen maken. Welke tussenstappen gemaakt moeten worden, moet uit de context afgeleid worden. Door hun geringere taalvaardigheid kunnen allochtone leerlingen meer moeite met complexe items hebben. Verder is het uiteraard mogelijk dat de context voor allochtone leerlingen minder herkenbaar is dan voor autochtone leerlingen. Dit geldt met name voor contextmateriaal dat cruciaal is voor het oplossen van het item.

7.4 Samenvatting

In de Verenigde Staten wordt relatief veel onderzoek gedaan naar itembias, maar men beperkt zich daarbij voornamelijk tot het opsporen van partijdige items (detectiefase). Het achterhalen van mogelijke oorzaken van itembias (verklaringsfase) heeft in Nederland maar ook in andere landen weinig aandacht gekregen. Goed gefundeerde taalkundig-inhoudelijke verklaringen inzake itembias voor allochtone leerlingen zijn niet beschikbaar. Omdat een theoretisch kader met betrekking tot bronnen van itembias voor allochtone leerlingen nog niet voorhanden is, hebben de conclusies die op basis van het onderhavige onderzoek hierover worden getrokken, een voorlopig karakter.

Bij het zoeken naar mogelijke oorzaken van itembias zijn niet alleen de medewerkers van het onderzoeksproject (van KUB en Cito) betrokken geweest

maar ook niet bij het project betrokken experts en leerlingen uit groep acht van het basisonderwijs.
De partijdige items in het voor- of nadeel van Turkse en/of Marokkaanse leerlingen zijn op grond van de in het geding zijnde vaardigheden eerst door elke projectmedewerkers afzonderlijk inhoudelijk geanalyseerd vanuit de vraag welke itemelementen mogelijk een bron van itembias vormen. Hierbij ging speciale aandacht uit naar elementen die voorkomen in items die partijdig zijn in het nadeel van beide groepen leerlingen en ontbreken in items die in het voordeel zijn van deze leerlingen en omgekeerd.
Bij deze inhoudsanalyse hebben de in 2.2 geformuleerde potentiële bronnen van itembias voor allochtone leerlingen als hypothesen gefungeerd. Vervolgens zijn de gemeenschappelijkheden in de analyses van de afzonderlijke project-medewerkers geïnventariseerd (7.1).
De inhoudelijke analyse leverde twee fundamentele problemen op. Enerzijds blijkt het moeilijk te zijn om met zekerheid aan te geven welk element de biasbron is en anderzijds blijkt dat bij sterk vergelijkbare items de ene keer wel sprake is van itembias en de andere keer niet. Tevens werd vastgesteld dat er een aanzienlijk aantal partijdige items is met onderlinge overeenkomsten.
De items die samen een cluster vormen, kunnen sterke aanwijzingen geven voor bronnen van itembias.
Hiermee komen we aan de zesde onderzoeksvraag van deze dissertatie:

6 *Welke bronnen van itembias voor allochtone leerlingen bevatten de opgaven van de Eindtoets Basisonderwijs 1987 en 1989?*

De resultaten van de inhoudsanalyse van items in een cluster die partijdig zijn volgens zowel de Mantel-Haenszel- als de IRT-procedure stemmen op een aantal punten overeen. Wanneer beide itembiasdetectieprocedures dezelfde bronnen van itembias aangeven, dan is de mate van zekerheid hierover groter. Deze overeenstemming geldt ten aanzien van de volgende bronnen van itembias:
- Tekstbegrip (begrijpend lezen): Items die naar globaal tekstbegrip vragen, hebben kans op itembias in het voordeel van Turkse en Marokkaanse leerlingen. Items die vragen naar de betekenis van een woord of een zin, door om een woordelijke of geparafraseerde herhaling van expliciet in de tekst gegeven informatie te vragen, hebben kans op itembias in het nadeel van deze leerlingen.
- Woordkennis en kennis van woordcombinaties: Items die vragen naar de betekenis van moeilijke woorden en waarvan de betekenis niet of moeilijk uit de context kan worden afgeleid, hebben een kans op itembias in het nadeel van deze leerlingen.
- Correct taalgebruik: Items die betrekking hebben op de kennis van de vorm van vaste woordcombinaties en conventies op het gebied van de zinsbouw, hebben een kans partijdig te zijn in het nadeel van Turkse en Marokkaanse leerlingen.
- Spelling: Items die vragen spelfouten in werkwoorden en in woorden met een vast woordbeeld aan te geven, hebben een kans op itembias in het voordeel van deze leerlingen.

De mate van zekerheid inzake bronnen van itembias is bij een deel van de clusters geringer, omdat de items alleen partijdig zijn volgens de Mantel-Haenszel-procedure. Dit geldt ook voor de items uit het cluster Referenties. Toch kan er hier sprake zijn een grotere mate van zekerheid inzake bronnen van itembias, omdat referenties bij twee clusters genoemd worden als een bron van itembias: Referenties en Rekenitems met relatief veel context.

Om meer aanwijzingen over bronnen van itembias te verkrijgen zijn ook niet bij het project betrokken experts gevraagd partijdige items in het voor- en nadeel van allochtone leerlingen op biasbronnen te beoordelen (7.2).
De experts blijken niet erg trefzeker te zijn in het opsporen van items die in het voor-, respectievelijk nadeel zijn van Turkse en Marokkaanse leerlingen. Er is vastgesteld hoe hoog de samenhang is tussen het aantal experts dat zegt dat een item moeilijker is voor allochtone leerlingen en de mate van itembias. De correlatie tussen het aantal experts dat zegt dat het item moeilijker is en de mate van itembias is niet hoog te noemen: $r = .30$ ($p<.01$).
De experts is niet gevraagd aan te geven bij welke items sprake is van itembias in het voor- of nadeel van allochtone leerlingen. Zij hebben aangegeven of een item moeilijker dan wel makkelijker voor allochtone leerlingen is dan voor autochtone leerlingen. Natuurlijk is de dimensie itembias in het voor- of nadeel niet gelijk aan de dimensie moeilijk-makkelijk, maar het is niet aannemelijk te veronderstellen dat dit verschil de geringe trefzekerheid van de experts verklaard.
De oordelen van de experts over de itemelementen die moeilijk zijn voor allochtone leerlingen blijken onderling in hoge mate te corresponderen en sluiten over het algemeen aan bij de bevindingen van de projectmedewerkers (zie 7.1.2). Daarnaast benadrukken de experts dat door de cultureel bepaalde voorkennis van allochtone leerlingen, die een rol kan spelen bij de items waarvoor contextmateriaal wordt gebruikt, sommige items voor deze leerlingen problematisch zijn. Allochtone leerlingen kunnen bij voorbeeld minder vertrouwd zijn met het onderwerp dat in een tekst aan de orde wordt gesteld. De experts zijn van oordeel dat veel opgaven geschreven zijn vanuit een Nederlandse cultuurkennis en daardoor veel allochtone leerlingen minder aanspreken. De experts maken geen onderscheid tussen moeilijke of makkelijke items voor Turkse dan wel voor Marokkaanse leerlingen.

Met een kleinschalig hardop-denken-experiment (7.3) is nagegaan hoe vaak allochtone en autochtone leerlingen bij partijdige items een fout antwoord geven door een itemelement dat voorlopig als bron voor itembias is aangewezen. Verder is onderzocht hoe vaak bij gemanipuleerde items door de itemmanipulatie een goed antwoord wordt gegeven. Gemanipuleerde items zijn items waarbij het itemelement dat als potentiële biasbron is aangewezen (bijvoorbeeld: 'Hoeveel moet hij betalen *inclusief* B.T.W.'), is vervangen door een itemelement waarvan verwacht wordt dat het geen bias veroorzaakt (bijvoorbeeld: 'Hoeveel moet hij betalen *met* B.T.W.'). Uit de Eindtoets Basisonderwijs zijn in totaal 17 items geselecteerd, die in de oorspronkelijke (partijdige) vorm zijn voorgelegd aan 11 allochtone en 11 autochtone leerlingen en die in gemanipuleerde vorm aan 11 andere allochtone en 11 andere autochtone leerlingen zijn voorgelegd. Het prestatieniveau van zowel de allochtone als de autochtone leerlingen die of de oorspronkelijke of de

gemanipuleerde versie maakten, was zoveel mogelijk vergelijkbaar.
De allochtone en autochtone leerlingen moesten eerst de hen voorgelegde oorspronkelijke, respectievelijk gemanipuleerde items goed bestuderen, daarna konden ze het naar hun oordeel goede antwoord aankruisen. Tot slot moesten ze zo uitgebreid en nauwkeurig mogelijk aangeven hoe ze aan hun antwoord gekomen waren. In de redeneringen van de leerlingen is getracht aanwijzingen te vinden omtrent bronnen van itembias.
Het hardop-denken-experiment geeft aanwijzingen dat biasbronnen voor een groot deel op het gebied van woordgebruik en impliciete zins- en tekst-verbanden gezocht moeten worden. Daarnaast lijken ongebruikelijke uitdrukkingen ('niet te laat') en een woordvormgelijkenis ('half procent'/'helft') tot problemen te leiden. Verder zijn grafische onduidelijkheden verwarrend.
De complexiteit van de items lijkt eveneens een rol te spelen. Complexe items vereisen doorgaans meer context en voor het oplossen van het item moet de leerling meestal een aantal tussenstappen maken. Welke tussenstappen gemaakt moeten worden, moet de leerling uit de context afleiden. Door hun geringere taalvaardigheid kunnen allochtone leerlingen meer moeite met complexe items hebben. Verder is het uiteraard mogelijk dat de context voor allochtone leerlingen minder herkenbaar is dan voor autochtone leerlingen. Dit is met name van belang bij contextmateriaal dat cruciaal is voor het oplossen van het item.
Het hardop-denken-experiment benadrukt op een aantal punten hetgeen eerder door de inhoudsanalyse van partijdige items al als biasbronnen naar voren is gekomen (7.1.2): woordkennis en kennis van woordcombinaties, impliciete zins- en tekstverbanden, grafische presentaties, rekenitems met relatief veel context.
De experts geven ook aan dat items over woordkennis en kennis van (idiomatische) woordcombinaties problematisch zijn. Zij voegen er aan toe dat de voorkennis van allochtone leerlingen minder kan aansluiten bij hetgeen in de items en teksten van de Eindtoets Basisonderwijs aan de orde wordt gesteld.

8 Samenvatting en discussie

8.1 Samenvatting van de Hoofdstukken 1-3

8.1.1 De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen (Hoofdstuk 1)

Om verschillen tussen schoolprestaties van allochtone en autochtone leerlingen te beschrijven worden veelvuldig toetsresultaten gebruikt. Tot nu toe is er echter nauwelijks onderzoek gedaan naar de vraag of veelgebruikte toetsen een geschikt middel zijn om de vaardigheid van allochtone leerlingen op het terrein van bepaalde onderwijsdoelstellingen te meten. Zowel onderzoekers als onderwijsgevenden blijken soms te twijfelen aan de bruikbaarheid van toetsen voor allochtone leerlingen.
Medewerkers van het Werkverband Taal en Minderheden van de Letterenfaculteit van de KUB en medewerkers van het project Eindtoets Basisonderwijs van het Cito besloten samen een onderzoeksproject uit te voeren waarin onderzoek gedaan wordt naar de bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen. De keuze om de Eindtoets Basisonderwijs te laten fungeren als object van onderzoek is vooral ingegeven door het feit dat er jaarlijks een groot aantal leerlingen aan de toets deelneemt (sinds 1992 meer dan 100 000). De toets heeft twee functies. Enerzijds verschaft de toets informatie over individuele leerlingen in verband met de overgang naar het voortgezet onderwijs, anderzijds levert de toets informatie voor de evaluatie van het onderwijsprogramma van de school. In het onderhavige onderzoek staat alleen de eerste functie van de Eindtoets Basisonderwijs centraal.
Nader onderzoek naar de bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen wordt eveneens ingegeven door de wens meer duidelijkheid te verkrijgen over het meten van vaardigheden bij een groep leerlingen waarvan de sociaal-culturele en linguïstische achtergrond over het algemeen sterk verschilt van die van de autochtone leerlingen. Bovendien is bekend dat allochtone leerlingen bij de meting van verschillende vaardigheden vaak lagere scores behalen. Met empirisch onderzoek is na te gaan of de scores van allochtone leerlingen op de Eindtoets Basisonderwijs een over- of een onderschatting dan wel een juiste weergave geven van de vaardigheid van deze leerlingen op de gemeten domeinen.

Het onderzoek richt zich op drie onderdelen. Ten eerste heeft het onderzoek betrekking op het beschrijven van trends in de schoolresultaten van allochtone en autochtone leerlingen. Met schoolresultaten worden hier toetsscores op (onderdelen van) de Eindtoets Basisonderwijs bedoeld en de gegevens van deze leerlingen omtrent de toelating tot en doorstroming in het voortgezet onderwijs. In de tweede plaats richt het onderzoek zich op toetsbias. In de onderhavige studie staat daarbij de vraag centraal hoe hoog de voorspellende waarde van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen is in vergelijking met de voorspellende waarde van het advies basisschool. Het derde onderdeel gaat over itembias.

In het onderzoek naar itembias zijn twee complementaire fasen onderscheiden: de detectie- en de verklaringsfase. In de eerste fase (detectiefase) zijn met statistische procedures items opgespoord waarbij sprake is van itembias. Bij een item is sprake van itembias wanneer leerlingen uit onderscheiden subgroepen (bijvoorbeeld allochtone en autochtone leerlingen) met dezelfde vaardigheid een ongelijke kans hebben om een bepaald item goed te beantwoorden. Voor onderzoek naar itembias, of anders gezegd naar partijdige items, worden over het algemeen twee soorten statistische procedures gebruikt: procedures die gebaseerd zijn op de klassieke testtheorie (bijvoorbeeld de Mantel-Haenszel-techniek) en op de itemresponsetheorie (IRT). Bij een procedure volgens het IRT-model wordt wel vastgesteld of de toetsitems de te meten vaardigheid adequaat representeren, bij een procedure op basis van de klassieke testtheorie gebeurt dit niet.
In de tweede fase van het onderzoek naar itembias (verklaringsfase) is een poging ondernomen om te onderzoeken wat bij een bepaald item de oorzaak van itembias zou kunnen zijn. Bij het door middel van inhoudsanalyse achterhalen van mogelijke oorzaken van itembias zijn niet alleen de project-medewerkers (van KUB en Cito) betrokken geweest, maar ook niet bij het onderzoeksproject betrokken experts en leerlingen uit groep acht van het basisonderwijs. Uit het onderzoek naar toets- en itembias kan blijken met welke aanpassingen de bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen eventueel vergroot kan worden.

8.1.2 Potentiële bronnen van toets- en itembias (Hoofdstuk 2)

De volgende potentiële determinanten van verschillen in de predictieve validiteit van toets (toetsbias) en advies basisschool zijn van belang.
Als een toets voor twee subgroepen dezelfde norm (regressievergelijking) hanteert om aan te geven welk type voortgezet onderwijs het beste gekozen kan worden, dan wordt het schoolsucces in het voortgezet onderwijs van de subgroep met de laagste toetsscore overschat, van de subgroep met de hoogste score onderschat. Omdat allochtone leerlingen meestal lagere toetsscores hebben dan autochtone leerlingen en omdat toetsen (ook de Eindtoets Basisonderwijs) meestal voor alle subgroepen dezelfde regressievergelijking hanteren om de positie op het extern criterium te schatten, mag verwacht worden, dat toetsen het schoolsucces van allochtone leerlingen overschatten en dat van autochtone leerlingen onderschatten.
Verder volgt uit onderzoek de verwachting dat het advies basisschool voor allochtone leerlingen eveneens een overschatting geeft van het schoolsucces in het voortgezet onderwijs.

Cummins (1984a) heeft een theoretisch raamwerk gepresenteerd dat door het accent op evaluatie van taalvaardigheid van allochtone leerlingen van belang kan zijn voor onderzoek naar itembias. Cummins' maakt onder meer onderscheid naar de mate waarin leerlingen, die een tweede taal (T2) leren, ondersteunende informatie ontvangen van de context waarin het taalgebruik is ingebed. Toetsitems verschillen in de mate waarin ze een beroep doen op contextmateriaal. In onderzoek naar oorzaken van itembias kan de aandacht uitgaan naar de omvang en naar de aard van het context-materiaal en naar de mate waarin de context allochtone leerlingen bij het beantwoorden items

ondersteunt.
Bij het opsporen van meer specifieke potentiële bronnen van itembias gaat de aandacht uit naar linguïstische elementen waarbij de prestaties van allochtone en autochtone leerlingen verschillen.
Potentiële bronnen van itembias op het niveau van woorden zijn:
- de betekenis van woorden;
- woorden met een lage woordfrequentie;
- woorden waarbij de context geen aanwijzingen geeft voor de betekenis van het woord;
- abstracte begrippen;
- ambigue woorden waarbij de context geen aanwijzingen geeft voor de betekenis van het woord.

Potentiële bronnen van itembias op het niveau van zinnen zijn:
- ontkennende zinnen;
- passieve zinnen;
- figuurlijk taalgebruik, specifieke idiomatische uitdrukkingen, metaforen.

Potentiële bronnen van itembias op het niveau van teksten zijn:
- teksten die een groot beroep doen op het geheugen;
- teksten waarvan de inhoud minder plausibel is;
- teksten met moeilijke referenties;
- teksten met ongebruikelijke of onjuiste aanduidingen over de structuur van de tekst.

Potentiële bronnen van itembias op het terrein van metalinguïstische vaardigheden zijn:
- items waarbij grammaticale onjuistheden opgespoord moeten worden;
- items waarbij aandacht voor en controle op het taalgebruik als zodanig een rol spelen.

Verder zijn er ook potentiële culturele bronnen van itembias. Te denken valt hierbij aan verschillen in voorkennis van de onderwerpen die in teksten aan de orde worden gesteld. Allochtone leerlingen kunnen door hun culturele achtergrond minder vertrouwd zijn met teksten die bijvoorbeeld qua inhoud gericht zijn op bekendheid met specifieke elementen van de Nederlandse samenleving.
Een andere potentiële culturele bron van itembias heeft betrekking op de verschillen tussen leerlingen in de mate waarin ze ervaring hebben in het maken van toetsen. Hierdoor kunnen sommige leerlingen minder weten welke taken ze moeten uitvoeren .
Verder moeten we er rekening mee houden dat allochtone leerlingen bepaalde elementen van het curriculum minder beheersen, omdat leerkrachten afhankelijk van het toekomstperspectief van de leerlingen meer of minder hoge eisen stellen aan de beheersing van bepaalde vaardigheden.

8.1.3 Beschrijving en verantwoording van de onderzoeksinstrumenten (Hoofdstuk 3)

Ten behoeve van het onderzoek naar toets- en itembias zijn vijf instrumenten ontwikkeld. De constructie van de Eindtoets Basisonderwijs en van de twee vragenlijsten voor het verzamelen van toelatings- en doorstroomgegevens in het voortgezet onderwijs behoren tot de cyclische activiteiten van het project Eindtoets Basisonderwijs van het Cito. Voor het verzamelen van achtergrond-

gegevens op leerling- en schoolniveau zijn speciaal voor dit onderzoek twee vragenlijsten ontwikkeld. Deze vragenlijsten zijn als Bijlage 1 en 2 integraal opgenomen. De onderzoekspopulaties bestaan uit de leerlingen van groep acht die in 1987, respectievelijk 1989 aan de Eindtoets Basisonderwijs deelnamen. De scholen is gevraagd in de periode van de toetsafname (halverwege februari 1987 en 1989) de vragenlijsten op leerling- en schoolniveau in te vullen. In mei na de toetsafname is aan de basisscholen gevraagd aan te geven naar welke school voor voortgezet onderwijs elke schoolverlater gaat. Ongeveer een jaar later is de betreffende scholen van voortgezet onderwijs verzocht te vermelden in welk type eerste leerjaar de leerling is geplaatst en naar welk type tweede leerjaar de leerling zal doorstromen of dat er sprake is van doubleren.

De Eindtoets Basisonderwijs is een schoolvorderingentoets die bestaat uit 180 vierkeuze-opgaven op het gebied van taal, rekenen en informatieverwerking. Deze drie toetsonderdelen worden elk weer onderverdeeld in opgaven-rubrieken. De toetsinhoud wordt verantwoord in het Doelenboek, de inhouds-verantwoording van de Eindtoets Basisonderwijs.
Op het leerlingrapport van de Eindtoets Basisonderwijs wordt naast de scores voor Taal, Rekenen en Informatieverwerking een standaardscore vermeld die door een equivaleringsprocedure van jaar tot jaar vergelijkbaar is. De toelatings- en doorstroomgegevens in het voortgezet onderwijs van een vorige generatie leerlingen worden gekoppeld aan deze standaardscore, waardoor voor de interpretatie van de toetsscores de relatie gelegd kan worden tussen de standaardscore die een leerling heeft behaald en de standaardscoreverdeling van de leerlingen die naar de onderscheiden typen voortgezet onderwijs zijn gegaan.

De vragenlijst op leerlingniveau bevat vragen over
- het land van herkomst;
- de schoolloopbaan van de leerling in het Nederlandse onderwijs;
- het oordeel van de leerkracht over het abstractieniveau van de leerling;
- het oordeel van de leerkracht over de afstand tussen het sociaal-culturele klimaat op school en thuis;
- het oordeel van de leerkracht over de geschiktheid van de leerling voor de verschillende typen van voortgezet onderwijs.

De vragenlijst op schoolniveau bevat vragen over
- de samenstelling van het sociaal milieu van de gehele schoolbevolking;
- het aantal leerlingen uit groep acht dat niet aan de Eindtoets Basisonderwijs deelneemt, omdat ze de Nederlandse taal onvoldoende beheersen om de opgaven te kunnen lezen.

Met de vragenlijsten van het toelatings- en doorstroomonderzoek is vastgesteld in welk type voortgezet onderwijs de leerling in het eerste en in het tweede leerjaar is geplaatst. Op de vragenlijst konden de leerkrachten uit het voortgezet onderwijs kiezen uit alle in de werkelijkheid voorkomende (combinaties van) typen voortgezet onderwijs.

8.2 Samenvatting van de Hoofdstukken 4 en 5 en discussie

8.2.1 Toetsresultaten en toelatings- en doorstroomgegevens van deelnemers aan de Eindtoets Basisonderwijs 1987 en 1989 (Hoofdstuk 4)

In hoofdstuk vier komen de eerste en tweede onderzoeksvraag aan de orde.

1 *Hoe ontwikkelen de Eindtoetsscores van allochtone en autochtone leerlingen zich van 1987 tot 1989?*

2 *Hoe verloopt de toelating en doorstroming van allochtone en autochtone leerlingen in het voortgezet onderwijs?*

De gemiddelde (geëquivaleerde) standaardscore van de onderscheiden etnische groepen op de Eindtoets Basisonderwijs 1987 en 1989 verschillen nauwelijks. De leerlingen van Marokkaanse en Turkse herkomst behalen in beide jaren de laagste gemiddelde standaardscore, gevolgd door de leerlingen van Surinaamse, Antilliaanse en Molukse herkomst. De Chinese leerlingen behalen vergeleken met alle andere etnische groepen (inclusief de autochtone leerlingen) de hoogste gemiddelde rekenscore. Dit wordt bevestigd door ander onderzoek in Nederland en de Verenigde Staten. De verschillen tussen de gemiddelde scores van autochtone en allochtone leerlingen zijn bij de spellingopgaven, met name bij de spelling van werkwoordvormen, relatief klein.

Wanneer we bij de toelating tot het voortgezet onderwijs naar de instroom van leerlingen van een vergelijkbaar prestatieniveau kijken, dan blijkt dat er meer autochtone dan allochtone leerlingen naar schooltypen met een lager gemiddeld prestatieniveau (IBO, LBO en LBO/AVO) gaan. Allochtone leerlingen hebben over het algemeen de overhand in het AVO in vergelijking met vergelijkbaar presterende autochtone leerlingen. Deze bevindingen zijn door andere onderzoekers herhaaldelijk bevestigd: allochtone leerlingen worden tot een hoger schooltype toegelaten dan op grond van toets- of testscores verwacht mag worden.

De relatieve voorsprong die allochtone leerlingen bij de start in het voortgezet onderwijs ten opzichte van autochtone leerlingen hebben, wordt volgens de gegevens van het onderhavige onderzoek voor een deel weer teniet gedaan aan het einde van het eerste leerjaar. Uit de doorstroomgegevens van leerlingen van een vergelijkbaar prestatieniveau blijkt dat allochtone leerlingen meer dan hun autochtone klasgenoten afstromen (blijven zitten of doorstromen naar een onderwijstype met een lager gemiddeld prestatieniveau). Ook blijkt dat allochtone leerlingen toegelaten tot LBO en MAVO, iets meer dan autochtone leerlingen doorstromen naar onderwijstypen met een hoger gemiddeld prestatieniveau. Opgemerkt moet worden dat de verschillen in af- en opstroom zowel in 1987 als in 1989 voorkomen, maar vrijwel nergens significant zijn. Er is tevens geconstateerd dat bij alle soorten scholengemeenschappen de verschillen tussen autochtone en allochtone leerlingen bij de doorstroming vanuit het eerste leerjaar niet significant zijn. Hoewel de afzonderlijke verschillen niet significant zijn, zijn er wel systematische verschillen: allochtone leerlingen gaan vanuit het eerste leerjaar van een scholengemeenschap zowel in

1987 als in 1989 doorgaans naar de schooltypen met een lager gemiddeld prestatieniveau.

8.2.2 Toetsbias in de Eindtoets Basisonderwijs 1987 en 1989 (Hoofdstuk 5)

In deze dissertatie wordt onderzoek naar toetsbias opgevat als het nagaan van de predictieve validiteit van de Eindtoets Basisonderwijs voor allochtone en autochtone leerlingen in vergelijking met die van het advies van de basisschool. De schaal voor schoolsucces, die in het onderzoek als afhankelijke variabele functioneert, wordt gedefinieerd door aan de onderwijsposities die de leerlingen in het voortgezet onderwijs innemen de waarde toe te kennen van de gemiddelde Cito-standaardscore van de leerlingen in dat onderwijstype.

In dit hoofdstuk komt de derde onderzoeksvraag aan bod.

3 *Hoe hoog is voor allochtone en autochtone leerlingen de voorspellende waarde van de Eindtoets Basisonderwijs in vergelijking met het advies van de basisschool.*

Om voor allochtone en autochtone leerlingen na te gaan hoe hoog de voorspellende waarde van de Eindtoets Basisonderwijs is in vergelijking met die van het advies basisschool, zijn eerst de produkt-moment correlatie-coëfficiënten tussen deze twee variabelen en de schaal voor schoolsucces berekend. Hieruit blijkt dat de voorspellende waarde van deze variabelen bij allochtone leerlingen doorgaans lager is dan bij autochtone leerlingen. Het voorspellen van schoolsucces voor allochtone leerlingen gaat dus minder trefzeker dan voor autochtone. Het advies basisschool verklaart meer variantie in schoolsucces dan de Cito-score. De Eindtoets voorspelt in 1987 het schoolsucces van allochtone en autochtone leerlingen even goed. In 1989 is de voorspellende waarde van de Eindtoets voor allochtone leerlingen iets lager dan voor autochtone leerlingen, maar dit geldt ook voor het advies basisschool.

Om de voorspellende waarde van het advies basisschool en de Eindtoets Basisonderwijs nauwkeuriger te kunnen analyseren is nagegaan hoe voor allochtone en autochtone leerlingen de regressielijnen van advies basisschool en Eindtoets Basisonderwijs op schoolsucces lopen. Uit de in het onderhavige onderzoek uitgevoerde analyses blijkt dat de regressielijnen van allochtone en autochtone leerlingen van de Eindtoets Basisonderwijs op de schaal voor schoolsucces significant verschillen ($p<.001$). Het verschil tussen de regressielijnen van allochtone en autochtone leerlingen van het advies basisschool op schoolsucces is in 1987 net ($p<.05$) en in 1989 niet significant. Het schoolsucces van allochtone leerlingen wordt door de Eindtoets Basisonderwijs overschat, dat van autochtone leerlingen onderschat. Als met de intercept en de helling bij de gemiddelde Eindtoetsscore van de populatie nagegaan wordt welke positie op de schaal voor schoolsucces geschat wordt, dan blijkt dat bij dezelfde Eindtoetsscore de positie van allochtone leerlingen op de schaal voor schoolsucces gemiddeld 0.19 standaarddeviatie hoger ligt dan die van autochtone leerlingen. Het advies basisschool blijkt het schoolsucces van allochtone en autochtone leerlingen met vrijwel dezelfde regressievergelijking te schatten. Deze resultaten corresponderen gedeeltelijk met de

veronderstellingen uit hoofdstuk twee. Zoals verwacht, overschat de Eindtoets Basisonderwijs de positie op de schaal voor schoolsucces van de subgroep met de laagste gemiddelde score (allochtone leerlingen). Tevens werd verwacht, dat het advies basisschool het schoolsucces van allochtone leerlingen zou overschatten, maar dit blijkt niet uit de regressielijnen van het advies basisschool op de schaal voor schoolsucces. In 8.2.3 wordt hier nader op ingegaan.

Om een beeld te krijgen van het causale effect van de onafhankelijke variabelen op de afhankelijke variabele schoolsucces zijn pad-analyses uitgevoerd. Uit de pad-analyses blijkt dat de schoolloopbaanmodellen zowel in 1987 als in 1989 meer variantie in schoolsucces verklaren bij autochtone dan bij allochtone leerlingen. Het effect van het advies basisschool op schoolsucces is in beide jaren groter dan het effect van de Cito-score. De Cito-score blijkt het schoolsucces van allochtone leerlingen beter te voorspellen dan het schoolsucces van autochtone leerlingen. Het leerjaar waarin de leerling is gestart in het Nederlandse basisonderwijs heeft, zoals verwacht, alleen een significant effect op schoolsucces bij allochtone leerlingen. Het effect van 'abstractievermogen' op schoolsucces is groot, zowel voor allochtone als voor autochtone leerlingen. Het effect van 'verschil school-thuis' op schoolsucces is voor allochtone leerlingen groter dan voor autochtone leerlingen.
Verder is gebleken dat het effect van de taalscore op schoolsucces bij autochtone leerlingen groter is dan bij allochtone. Bij allochtone leerlingen is het effect van de taal- en rekenscore vrijwel gelijk. Het effect van de informatieverwerkingscore is bij allochtone leerlingen geringer.
Er zijn ook pad-analyses per etnische groep uitgevoerd. Bij autochtone leerlingen en bij leerlingen met ouders uit Zuid-Europa en Marokko blijkt dat zowel in 1987 als in 1989 het effect van het advies basisschool dominant is. Bij de Surinaams-Creoolse leerlingen is het effect van het advies basisschool en de Cito-score in beide jaren vrijwel even groot. Verder valt op dat het effect van de informatieverwerkingscore op schoolsucces alleen bij autochtone, Turkse en Marokkaanse leerlingen significant is. Bij de Chinese leerlingen heeft in 1987 en 1989 alleen de rekenscore effect op schoolsucces. De Chinese leerlingen behalen de hoogste rekenscores en deze scores overtreffen in beide jaren de taal- en informatieverwerkingscores met betrekking tot het effect op schoolsucces.

8.2.3 Discussie

Het interpreteren van de verschillen tussen de regressielijnen als kwaliteitskenmerk van advies basisschool of toetsscore is een moeilijke zaak, omdat hierbij rekening gehouden moet worden met de complexiteit van de onderwijspraktijk. Het advies basisschool en de score van de Eindtoets Basisonderwijs en de onderwijspositie die leerlingen na een jaar in het voortgezet onderwijs innemen, zijn geen onafhankelijke factoren in het schoolkeuzeproces. Bij de overgang naar het voortgezet onderwijs brengt de directeur basisschool een advies uit aan de ouders en aan de toelatingscommissie van het voortgezet onderwijs. Ouders en leerling beschikken over het advies basisschool en veelal over de Eindtoetsscore wanneer zij hun wens inzake de schoolkeuze definitief bepalen. De ouders kunnen hun kind, gegeven een bepaald prestatieniveau, in

de meeste gevallen bij verschillende, concurrerende scholen aanmelden. Zo kan een kind met bijvoorbeeld een MAVO-advies vaak kiezen voor een categoriale MAVO of voor een MAVO/HAVO/VWO, al of niet van dezelfde denominatie. De toelatingscommissie van de school waarbij het kind uiteindelijk is aangemeld, beoordeelt het advies basisschool, de Eindtoetsscore en de wens van ouders/kind bij het nemen van een beslissing over de toelating.

De invloed van advies en toetsscore op de schoolkeuze en het schoolsucces na een jaar voortgezet onderwijs zou wellicht te verwaarlozen zijn op het volgende denkbeeldige eiland. Op dit eiland zijn van de leerlingen aan het einde van het basisonderwijs een toetsscore en een advies basisschool beschikbaar. Alle schoolverlaters gaan op dit eiland naar een zo breed mogelijke scholengemeenschap met een driejarige brugperiode waarin zowel de potentiële IBO- als de potentiële VWO-abituriënten bij elkaar in de klas zitten. De docenten op deze scholengemeenschap zijn niet bekend met het advies en de toetsscore van de basisschool. De bevordering aan het einde van het eerste leerjaar is alleen gebaseerd op het prestatieniveau van de leerlingen in de onderwezen vakken. Op dit denkbeeldige eiland kan de regressie van advies en toetsscore op schoolsucces beoordeeld worden als een kwaliteitskenmerk van de voorspellers, omdat het advies basisschool en de toetsscore op zich hier geen invloed kunnen uitoefenen op het schoolsucces in het voortgezet onderwijs.

Voor de beoordeling van de predictieve validiteit van advies basisschool en Eindtoetsscore als kwaliteitskenmerk moeten we met de praktijk van het schoolkeuzeproces rekening houden. De hogere correlatie tussen advies basisschool en schoolsucces in vergelijking met de correlatie tussen toetsscore en schoolsucces kan voor een deel ook verklaard worden uit de grotere invloed die het advies basisschool op de schoolkeuze heeft, hetgeen een aannemelijke veronderstelling lijkt te zijn (vgl. Van Essen, 1983; Cito, 1986b). Het is niet denkbeeldig dat onderzoek naar de predictieve validiteit van advies basisschool en Eindtoets Basisonderwijs deels bestaat uit het beantwoorden van de vraag wat de grootste invloed heeft in het schoolkeuzeproces. Dit gevaar is met name aanwezig wanneer de maat voor schoolsucces betrekking heeft op een relatief korte schoolloopbaan.
Om meer inzicht in de verschillen in de schoolloopbaan van allochtone en autochtone leerlingen uit 1987 en 1989 te krijgen zijn de verschillen tussen deze leerlingen op verschillende momenten geanalyseerd.

De basisscholen hebben halverwege februari (1987 en 1989) van elke Eindtoetsdeelnemer het schoolkeuze-advies verstrekt. Het verschil tussen het advies basisschool van allochtone en autochtone was over beide jaren gerekend gemiddeld 0.33 standaarddeviatie. De Eindtoetsscore was begin maart beschikbaar. Het verschil tussen beide subgroepen op de Eindtoets Basisonderwijs was gemiddeld 0.62 standaarddeviatie. Het verschil tussen allochtone en autochtone leerlingen is op de schaal voor schoolsucces in het eerste leerjaar gemiddeld 0.26 standaarddeviatie en in het tweede leerjaar 0.28 standaarddeviatie. Een samenvatting van deze gegevens wordt in tabel 8.1 gegeven. De verschillen tussen beide groepen leerlingen zijn op alle momenten significant ($p<.001$).

Tabel 8.1 Gemiddelde verschillen tussen allochtone en autochtone leerlingen op verschillende momenten (in standaarddeviaties)

Moment	Bron	Verschil
februari groep acht	advies basisonderwijs	0.33
maart groep acht	Eindtoets Basisonderwijs	0.62
klas 1 voortg. ond.	schaal voor schoolsucces klas 1	0.26
klas 2 voortg. ond.	schaal voor schoolsucces klas 2	0.28

Uit tabel 8.1 blijkt dat de afstand tussen beide groepen leerlingen bij de Eindtoets Basisonderwijs bijna twee keer zo groot is als bij het advies basisschool. In het eerste leerjaar van het voortgezet onderwijs zijn de verschillen tussen beide subgroepen het kleinst. Uit onderzoek weten we dat de directeur basisschool allochtone leerlingen gemiddeld hogere adviezen geeft, dan op grond van de toetsscores zou worden verwacht (Mulder & Tesser, 1991; Driessen, 1991a; Van Langen & Jungbluth, 1992; Meijnen & Riemersma, 1992). Bovendien komen deze leerlingen voor een zeer belangrijk deel ook in de geadviseerde schooltypen terecht of nog hoger. Gezien de regressielijnen van advies basisschool op schoolsucces lijkt het advies een zeer grote invloed te hebben op de positie in het voortgezet onderwijs.
De relatief kleine verschillen tussen allochtone en autochtone leerlingen in het eerste leerjaar moeten waarschijnlijk mede toegeschreven worden aan de invloed die de ouders/het kind en de toelatingscommissie in het schoolkeuze-proces hebben. Het is denkbaar dat de ouders van allochtone leerlingen aandringen op een hogere schoolkeuze dan het advies aangeeft (vgl. Mulder & Tesser, 1991; Van Langen & Jungbluth, 1992, De Wit, Suhre & Mulder, 1993). Een eventueel lage score op de Eindtoets Basisonderwijs wordt wellicht minder relevant geacht. Ook kan gezegd worden dat de toelatingscommissies de wensen van de ouders van allochtone leerlingen in deze veelal niet in de weg staan.

Aan het begin van het tweede leerjaar zijn de verschillen tussen beide subgroepen weer groter dan in het eerste leerjaar. Dit wordt wellicht veroorzaakt door het feit dat allochtone leerlingen bij de overgang naar het tweede leerjaar minder succesvol zijn dan autochtone leerlingen. De afstroom van allochtone leerlingen – zo bleek in hoofdstuk vier – is 1.3 tot 3.8 keer zo hoog (blijven zitten of doorstroming naar een lager onderwijsniveau). Ook Mulder & Tesser (1991) rapporteren dat de afstroom van allochtone leerlingen aan het einde van het eerste leerjaar groter is dan die van autochtone leerlingen. De Wit, Suhre & Mulder (1993) onderzochten hoe de toelating en doorstroming is verlopen van allochtone en autochtone leerlingen die in het derde leerjaar van het voortgezet onderwijs verblijven. Zij maakten hierbij onderscheid tussen

leerlingen die het schoolkeuze-advies van de basisschool hebben opgevolgd, tussen leerlingen die een lager schooltype en tussen leerlingen die een hoger schooltype hebben gekozen dan het advies basisschool. Het advies blijkt, na controle voor schoolkeuze, vrijwel geen effect meer te hebben op de positie in het derde leerjaar voortgezet onderwijs. De Wit, Suhre & Mulder (1993) concluderen dat er geen nadelige effecten zijn van de relatief hoge adviezen aan allochtone leerlingen. Zij merken op dat de relatief grote afstroom van allochtone leerlingen in het voortgezet onderwijs veel meer een gevolg is van het afwijken van het advies basisschool, dan van het advies zelf.

Uit het onderhavige onderzoek blijkt dat bij allochtone leerlingen de keuze van een school voor voortgezet onderwijs minder trefzeker gebeurt dan bij autochtone leerlingen. De betrokkenen in het schoolkeuzeproces van allochtone leerlingen moeten beseffen dat het advies basisschool een belangrijke graadmeter is voor de schoolkeuze. Uit hoofdstuk vijf is gebleken dat de Eindtoets Basisonderwijs voor allochtone leerlingen niet minder bruikbaar is dan voor autochtone leerlingen. Voor het maken van de schoolkeuze van allochtone leerlingen kunnen in de huidige situatie over het algemeen de Eindtoetsscores op Taal en Rekenen een even sterk accent krijgen, terwijl de score op Informatieverwerking in de meeste gevallen minder benadrukt kan worden.

De directeur basisschool geeft over het algemeen allochtone leerlingen hogere adviezen dan op grond van toetsscores verwacht mag worden. Bovendien komen deze leerlingen voor een belangrijk deel ook in de geadviseerde schooltypen terecht of nog hoger. De positie van allochtone leerlingen in de beginfase van het voortgezet onderwijs is nog wel een punt van aandacht. Omdat het voortgezet onderwijs na een jaar al niet bij machte is om de toegelaten allochtone leerlingen eenzelfde start te geven als de autochtone leerlingen, kan opgemerkt worden dat het wenselijk zou zijn als het voortgezet onderwijs zijn verantwoordelijkheid voor de toegelaten allochtone leerlingen wat meer zou kunnen effectueren.

Er is reeds gesteld dat de regressie van advies en toetsscore op schoolsucces van allochtone en autochtone leerlingen niet goed te interpreteren is als kwaliteitskenmerk van de Eindtoets Basisonderwijs of het advies basisschool. Er zijn minstens twee mogelijkheden om de regressie van voorspeller op schoolsucces meer als kwaliteitskenmerk te kunnen hanteren.

- Ten eerste zou een maat voor schoolsucces betrekking moeten hebben op een langere schoolloopbaan in het voortgezet onderwijs. Aan het einde van twee of drie jaar voortgezet onderwijs is immers verder uitgekristalliseerd wat meer of minder succesvolle leerlingen zijn.
- Ten tweede zou een maat voor schoolsucces beschikbaar moeten zijn die bestaat uit een schoolvorderingentoets voor de kernvakken van het voortgezet onderwijs. Deze toets zou bijvoorbeeld afgenomen kunnen worden aan het einde van het eerste leerjaar voortgezet onderwijs. Bij kernvakken kan gedacht worden aan de vakken Nederlands, Engels en wiskunde. Als deze toets weinig of geen items zou bevatten die partijdig zijn voor allochtone leerlingen, is het wellicht mogelijk de predictieve validiteit van het advies basisschool en de Eindtoets Basisonderwijs voor allochtone leerlingen

nauwkeuriger te beoordelen. Bovendien zou in het algemeen gesproken een bruikbaar instrument voor schoolloopbaanonderzoek beschikbaar zijn.

8.3 Samenvatting van de Hoofdstukken 6 en 7 en discussie

In het onderzoek naar itembias zijn twee elkaar aanvullende fasen onderscheiden. In de eerste fase worden met statistische procedures partijdige items opgespoord. De eerste fase, de detectiefase, komt in hoofdstuk zes aan de orde. In de tweede fase wordt ingegaan op de vraag wat bij een bepaald item de oorzaak van itembias zou kunnen zijn. De tweede fase, de verklaringsfase, wordt in hoofdstuk zeven beschreven.

8.3.1 Itembias in de Eindtoets Basisonderwijs 1987 en 1989 (Hoofdstuk 6)

Een item is partijdig wanneer leerlingen uit onderscheiden subgroepen, maar met hetzelfde vaardigheidsniveau een ongelijke kans hebben om het betreffende item goed te beantwoorden. Voor het opsporen van partijdige items in de Eindtoets Basisonderwijs 1987 en 1989 voor allochtone leerlingen zijn twee itembiasdetectieprocedures gebruikt. Het computerprogramma One Parameter Logistic Model (OPLM) van Verhelst (1992) is gebruikt als procedure die gebaseerd is op de itemresponsetheorie (IRT); het Mantel-Haenszel-programma (Verhelst, 1988) is gehanteerd als procedure gebaseerd op de klassieke testtheorie. Het Mantel-Haenszel-programma gaat van de assumptie uit dat het totaal aantal goed gemaakte opgaven een juiste schatting is van de te meten vaardigheid. Onder het IRT-model wordt getoetst of deze aanname juist is door te onderzoeken of de items een eendimensionele schaal vormen.
In verband met het aantal waarnemingen per etnische groep zijn de items van de drie toetsonderdelen Taal, Rekenen en Informatieverwerking uit de Eindtoets Basisonderwijs 1987 en 1989 alleen voor Turkse en Marokkaanse in vergelijking met autochtone leerlingen op itembias onderzocht. De items van Eindtoets Basisonderwijs 1987 en 1989 zijn eerst met de Mantel-Haenszel-procedure onderzocht. Daarna is met OPLM vastgesteld welke items een eendimensionele schaal vormen en als basis kunnen dienen voor onderzoek naar itembias onder het IRT-model. De items van elke eendimensionele schaal zijn met OPLM op itembias onderzocht en ook nog met de Mantel-Haenszel-procedure. Hierdoor is het mogelijk de resultaten van de Mantel-Haenszel-procedure met die van de IRT-procedure te vergelijken.

Uit de resultaten van de analyses naar itembias blijkt dat het moeilijk is om aan te geven in welke mate de Eindtoets Basisonderwijs 1987 en 1989 partijdige items bevatten. De verschillende analyses laten een wisselend beeld zien.
Bij één van de Mantel-Haenszel-analyses is gebleken dat voor Turkse en Marokkaanse leerlingen zowel in 1987 als in 1989 ongeveer de helft van de 60 taalitems partijdig is, terwijl de 60 rekenitems voor éénderde tot éénvijfde deel uit partijdige items bestaat en de 60 informatieverwerkingitems voor éénderde deel. De resultaten van deze Mantel-Haenszel-analyses laten slechts voorlopige conclusies toe, omdat nog niet vastgesteld is of de items van een toetsonderdeel een eendimensionele schaal vormen, waardoor het niet

uitgesloten is dat een aantal items partijdig is, omdat ze zowel voor autochtone als voor allochtone leerlingen multidimensioneel zijn. Bij de analyses met de IRT-procedure is gebleken dat het aantal partijdige items voor Turkse en/of Marokkaanse leerlingen beduidend geringer is: 20 van de in totaal 360 geanalyseerde items (=6%). Opgemerkt moet worden dat bij de IRT-analyses de totaalscore waarmee de leerlingen ingedeeld worden in groepen met een vergelijkbaar vaardigheidsniveau gebaseerd is op de items van de eendimensionele schaal, bij de Mantel-Haenszel-analyse op de totaalscore van het betreffende toetsonderdeel. Als we de items van de eendimensionele schaal gebruiken voor de totaalscore bij de Mantel-Haenszel-analyse dan blijken er 45 van de 360 geanalyseerde items partijdig te zijn (=13%). In totaal zijn er 13 items partijdig bij zowel de IRT- als de Mantel-Haenszel-procedure (=4%). Items zijn nooit in het voordeel voor Turkse leerlingen en tegelijkertijd in het nadeel van Marokkaanse leerlingen of omgekeerd.
Uit nadere analyse blijkt dat beide procedures (met als totaalscore de items van de eendimensionele schaal) in 87% van de gevallen overeenstemmen in het detecteren van (on)partijdige items. Dit beeld komt overeen met de resultaten van andere onderzoekers. De stabiliteit van de IRT- en de Mantel-Haenszel-procedure is vrijwel gelijk: in twee steekproeven wijst de Mantel-Haenszel-procedure bij 86% van de items en de IRT-procedure bij 89% van de items in beide steekproeven een item als partijdig aan. Bij vergelijking van de resultaten van de verschillende Mantel-Haenszel-analyses blijkt dat het veel uitmaakt op welke items de totaalscore wordt gebaseerd.

Sommige items zijn bij alle analyses partijdig, andere items zijn dit nooit en enkele items zijn bij een deel van de analyses partijdig. Voordat gestart kan worden met de inhoudelijke analyse van partijdige items uit de Eindtoets Basisonderwijs 1987 en 1989 moet echter wel vastgesteld worden welke items partijdig zijn en welke niet. Hiermee komen we aan de vierde en vijfde onderzoeksvraag:

4 *Welke statistische procedure verdient de voorkeur voor het opsporen van itembias bij de Eindtoets Basisonderwijs?*

5 *Welke opgaven zijn voor allochtone leerlingen significant moeilijker of makkelijker dan voor autochtone leerlingen met een vergelijkbaar prestatieniveau?*

De verschillende resultaten worden met name beïnvloed door het feit dat de Mantel-Haenszel-procedure, in tegenstelling tot de IRT-techniek, gebaseerd is op de aanname dat het totaal aantal goed gemaakte opgaven een adequate schatting is van de te meten vaardigheid. Bij de IRT-procedure wordt deze assumptie getoetst, waardoor we er bij deze laatste procedure meer vanuit kunnen gaan dat we itembiasonderzoek doen met leerlingen van hetzelfde vaardigheidsniveau. Omdat bij onderzoek naar itembias leerlingen juist op die vaardigheid gematcht moeten worden, gaat de voorkeur uit naar een itembias-detectietechniek die gebaseerd is op het IRT-model.

Hoewel in theoretisch opzicht de voorkeur uitgaat naar IRT-procedures, verdient het feit, dat de Eindtoets Basisonderwijs samengesteld en geanalyseerd

wordt volgens de klassieke testtheorie, nog nadrukkelijk aandacht. In de jaren 1987 en 1989 is bij de psychometrische analyse en rapportage van de Eindtoets Basisonderwijs de klassieke testtheorie gehanteerd. In die jaren zijn de scores van de deelnemers op de toetsonderdelen en het totaal van de Eindtoets Basisonderwijs gebaseerd op het ongewogen aantal goed gemaakte opgaven. Dit impliceert een arbitraire weging van het item en van de vaardigheidsdimensie. Wanneer bij de rapportage van de Eindtoets Basisonderwijs 1987 en 1989 het eenparameter model zou zijn gehanteerd en weging van de schalen zou zijn toegepast, dan zou dit voor allochtone en autochtone leerlingen tot andere totaalscores kunnen leiden. Omdat de correlaties tussen gewogen en ongewogen scores, respectievelijk schalen doorgaans hoog zijn te noemen, zouden de totaalscores waarschijnlijk niet dramatisch verschillen van de in 1987 en 1989 aan de toetsdeelnemers gerapporteerde scores, maar het betekent wel dat we bij de keuze van de itembiasdetectieprocedure niet alleen vanuit theoretische voorkeuren kunnen vertrekken.
Gezien het bovenstaande ligt het voor de hand om voor het detecteren van partijdige items zowel een klassieke testtheorie- als een IRT-procedure te kiezen. De items die volgens de klassieke testtheorie (Mantel-Haenszel) en de items die volgens de IRT-procedure partijdig zijn, zijn inhoudelijk geanalyseerd en er is nagegaan in hoeverre de resultaten van de inhoudelijke analyses van de beide soorten partijdige items overeenstemmen. De mate van overeenstemming geeft een indicatie voor de graad van zekerheid waarmee we inhoudelijke elementen van items als bron van itembias kunnen aanmerken.

8.3.2 Bronnen van itembias (Hoofdstuk 7)

In de Verenigde Staten doet men relatief veel onderzoek naar itembias, maar men beperkt zich daarbij voornamelijk tot de detectiefase. Met het zoeken naar mogelijke oorzaken van itembias (verklaringsfase) is over het algemeen nog weinig ervaring opgedaan. Goed gefundeerde taalkundig-inhoudelijke verklaringen met betrekking tot itembias voor allochtone leerlingen ontbreken geheel. Omdat een theoretisch kader inzake mogelijke oorzaken van itembias voor allochtone leerlingen niet beschikbaar is, hebben de conclusies die op basis van het onderhavige onderzoek hierover worden getrokken, een voorlopig karakter.

Aan het zoeken naar potentiële bronnen van itembias hebben drie groepen personen deelgenomen, respectievelijk de medewerkers van het onderzoeksproject (van KUB en Cito), niet bij het project betrokken experts en leerlingen uit groep acht van het basisonderwijs.
De items die in het voor- of nadeel van Turkse en/of Marokkaanse leerlingen partijdig zijn, zijn op grond van de te meten vaardigheid eerst door de projectmedewerkers onafhankelijk van elkaar inhoudelijk geanalyseerd met de vraag welke itemelementen mogelijk de bron van itembias vormen. Hierbij was bijzondere aandacht gevestigd op elementen die voorkomen in items die partijdig zijn in het nadeel van beide groepen leerlingen en ontbreken in items die in het voordeel zijn van deze leerlingen en omgekeerd. Bij deze inhoudsanalyse hebben de in hoofdstuk twee geformuleerde potentiële bronnen van itembias voor allochtone leerlingen als hypothesen gefungeerd. Vervolgens zijn de gemeenschappelijkheden in de analyses van de afzonderlijke project-

medewerkers geïnventariseerd.
De inhoudelijke analyse leverde twee fundamentele problemen op. Enerzijds blijkt het moeilijk te zijn om met zekerheid aan te geven welk itemelement de biasbron is en anderzijds blijkt dat bij sterk vergelijkbare items de ene keer wel sprake is van itembias en de andere keer niet. Tevens werd vastgesteld dat er een aanzienlijk aantal partijdige items is met onderlinge overeenkomsten.
De items die samen een cluster vormen, kunnen sterke aanwijzingen geven voor bronnen van itembias.
Hiermee komen we aan de zesde onderzoeksvraag:

6 *Welke bronnen van itembias voor allochtone leerlingen bevatten de opgaven van de Eindtoets Basisonderwijs 1987 en 1989?*

De resultaten van de inhoudsanalyse van items in een cluster die partijdig zijn volgens zowel de Mantel-Haenszel- als de IRT-procedure stemmen op een aantal punten overeen. Wanneer beide itembiasdetectieprocedures dezelfde bronnen van itembias aangeven, dan is de mate van zekerheid hierover groter. Deze overeenstemming geldt ten aanzien van de volgende bronnen van itembias (Als achter een bron van itembias 'zie hoofdstuk twee' staat, wordt daarmee bedoeld dat de betreffende biasbron in hoofdstuk twee wordt genoemd als potentiële biasbron).
- Tekstbegrip (begrijpend lezen): Items die naar globaal tekstbegrip vragen, hebben kans op itembias in het voordeel van Turkse en Marokkaanse leerlingen. Items die vragen naar de betekenis van een woord of een zin, door om een woordelijke of geparafraseerde herhaling van expliciet in de tekst gegeven informatie te vragen, hebben kans op itembias in het nadeel van deze leerlingen (zie hoofdstuk twee).
- Woordkennis en kennis van woordcombinaties: Items die vragen naar de betekenis van moeilijke woorden en waarvan de betekenis niet of moeilijk uit de context kan worden afgeleid, hebben een kans op itembias in het nadeel van deze leerlingen (zie hoofdstuk twee).
- Correct taalgebruik: Items die betrekking hebben op de kennis van de vorm van vaste woordcombinaties en conventies op het gebied van de zinsbouw, hebben een kans partijdig te zijn in het nadeel van Turkse en Marokkaanse leerlingen (zie hoofdstuk twee).
- Spelling: Items die vragen spelfouten in werkwoorden en in woorden met een vast woordbeeld aan te geven, hebben een kans op itembias in het voordeel van deze leerlingen.

De mate van zekerheid inzake bronnen van itembias is bij een deel van de clusters geringer, omdat de items alleen partijdig zijn volgens de Mantel-Haenszel-procedure. Dit geldt ook voor de items uit het cluster Referenties. Toch kan er hier sprake zijn een grotere mate van zekerheid inzake bronnen van itembias, omdat referenties bij twee clusters genoemd worden als een bron van itembias: Referenties en Rekenitems met relatief veel context (zie hoofdstuk twee).

Om meer aanwijzingen over bronnen van itembias te verkrijgen zijn niet bij het project betrokken experts gevraagd partijdige items in het voor- en nadeel van allochtone leerlingen op biasbronnen te beoordelen. De experts blijken niet erg

trefzeker te zijn in het opsporen van items die in het voor-, respectievelijk nadeel zijn van Turkse en Marokkaanse leerlingen. Er is nagegaan hoe hoog de samenhang is tussen het aantal experts dat zegt dat een item moeilijker is voor allochtone leerlingen en de mate van itembias. De correlatie tussen het aantal experts dat zegt dat het item moeilijker is en de mate van itembias is niet hoog te noemen: $r = .30$ ($p<.01$). De oordelen van de experts over de itemelementen die moeilijk zijn voor allochtone leerlingen blijken onderling in hoge mate te corresponderen en sluiten over het algemeen aan bij de bevindingen van de projectmedewerkers. Daarnaast benadrukken de experts dat door de cultureel bepaalde voorkennis van allochtone leerlingen, die een rol kan spelen bij de items waarvoor contextmateriaal wordt gebruikt, sommige items voor deze leerlingen problematisch zijn (zie hoofdstuk twee). Zo kunnen allochtone leerlingen minder vertrouwd zijn met het onderwerp dat in een tekst aan de orde wordt gesteld. De experts zijn van oordeel dat veel opgaven geschreven zijn vanuit een Nederlandse cultuurkennis en daardoor veel allochtone leerlingen minder aanspreken. De experts maken geen onderscheid tussen moeilijke of makkelijke items voor Turkse dan wel voor Marokkaanse leerlingen.

Door middel van een kleinschalig hardop-denken-experiment is onderzocht hoe vaak allochtone en autochtone leerlingen partijdige items fout beantwoorden door een element in het item dat waarschijnlijk de itembias veroorzaakt. Verder is onderzocht hoe vaak bij gemanipuleerde items door de itemmanipulatie een goed antwoord wordt gegeven. Gemanipuleerde items zijn items waarbij het itemelement dat vermoedelijk de biasbron is (bijvoorbeeld: 'Hoeveel moet hij betalen *inclusief* B.T.W.'), is vervangen door een itemelement waarvan verwacht wordt dat het geen bias veroorzaakt (bijvoorbeeld: 'Hoeveel moet hij betalen *met* B.T.W.'). De allochtone en autochtone leerlingen moesten eerst de hen voorgelegde oorspronkelijke, respectievelijk gemanipuleerde items goed bestuderen, daarna konden ze het naar hun oordeel goede antwoord aankruisen. Tot slot moesten ze zo uitgebreid en nauwkeurig mogelijk aangeven hoe ze aan hun antwoord gekomen waren. In de redeneringen van de leerlingen is getracht aanwijzingen te vinden omtrent bronnen van itembias.
Het hardop-denken-experiment geeft aanwijzingen dat biasbronnen voor een groot deel op het gebied van woordgebruik en impliciete zins- en tekst-verbanden gezocht moeten worden (zie hoofdstuk twee). Verder zijn grafische onduidelijkheden verwarrend. De complexiteit van de items lijkt eveneens een rol te spelen. Complexe items vereisen doorgaans meer context en voor het oplossen van het item moet de leerling meestal een aantal tussenstappen maken. Welke tussenstappen gemaakt moeten worden, moet de leerling uit de context afleiden. Door hun geringere taalvaardigheid kunnen allochtone leerlingen meer moeite met complexe items hebben. Verder is het uiteraard mogelijk dat de context voor allochtone leerlingen minder herkenbaar is dan voor autochtone leerlingen. Dit is met name van belang bij contextmateriaal dat cruciaal is voor het oplossen van het item (zie hoofdstuk twee).

8.3.3 Discussie

Concluderend kunnen we zeggen dat onderzoek naar itembias met veel onzekerheden is omgeven.
In de eerste plaats kan niet duidelijk worden aangegeven of een item partijdig is of niet (Hoofdstuk 6). Dit wordt deels veroorzaakt door het ontbreken van volledige overeenstemming tussen de resultaten van de verschillende itembias-detectieprocedures. De onzekerheid wordt voor het grootste deel veroorzaakt door de verschillen die ontstaan door de keuze voor het ene of andere criterium om de leerlingen in te delen in niveaugroepen (een- of multidimensionele totaalscore). Het is in dit soort onderzoek gebruikelijk om meer dan één analysetechniek te hanteren en om meer dan één criterium als totaalscore te gebruiken om leerlingen in niveaugroepen in te delen. Door het hanteren van verschillende procedures ontstaan er gradaties in de partijdigheid van items.
In de tweede plaats kan niet duidelijk aangegeven worden welk element van een item verantwoordelijk is voor de itembias (Hoofdstuk 7). De bron van itembias kan bijvoorbeeld in het contextmateriaal, in de vraag, de antwoord-mogelijkheden en in de te meten (deel)vaardigheid zitten. Naarmate het contextmateriaal omvangrijker wordt, is het moeilijker om met zekerheid de biasbron aan te wijzen. Zo kan de voorkennis omtrent hetgeen in teksten aan de orde wordt gesteld van invloed zijn op de itemantwoorden.
Scheuneman (1985), Uiterwijk & Vallen (1991) en Schmitt e.a. (1992) geven aan dat via verschillende wegen (inhoudsanalyse, expertbevraging, experimenten) informatie over bronnen van itembias verzameld kan worden. Als uiteindelijk – zoals in de onderhavige studie – met een zekere mate van waarschijnlijkheid een aantal biasbronnen opgespoord zijn, dan blijft nog de vraag over of de biasbronnen afbreuk doen aan de constructvaliditeit van de toets of niet.

Hiermee komen we aan bij de slotvraag.
Welke richtlijnen zijn er aan toetsconstructeurs te geven om itembias voor allochtone leerlingen te vermijden?
Wanneer een bron van itembias behoort tot de te meten vaardigheid, dan meet het item wat het beoogt te meten. Wanneer een biasbron niet tot de te meten vaardigheid behoort dan doet het item afbreuk aan de constructvaliditeit van de toets. Zo moeten bronnen van itembias op het gebied van woordenschat geen rol spelen bij rekenitems, ze mogen wel opgenomen zijn in taalitems in zoverre die items een bijdrage leveren aan het meten van woordkennis. Itembias als zodanig beperkt de constructvaliditeit van een toets dus niet in alle gevallen. Uit hoofdstuk zeven blijkt dat het aantal bronnen van itembias dat met een grote mate van zekerheid bias veroorzaakt, in feite gering is. Het detecteren van partijdige items en het verklaren van de oorzaken ervan is een lange weg met een bescheiden opbrengst. Informatie over potentiële bronnen van itembias komt ook beschikbaar uit elk empirisch onderzoek dat duidelijk maakt bij welke kennis- en vaardigheidsaspecten allochtone leerlingen significant lager scoren dan autochtone leerlingen. Wanneer de toetsconstructeur over gedetailleerde informatie op dit gebied beschikt, kan hij voorkomen dat hij inzake allochtone leerlingen items construeert die afbreuk doen aan de constructvaliditeit van een toets. Hiervoor is het van belang dat in empirisch onderzoek kennis- en vaardigheidsaspecten op een zo gedetailleerd mogelijk niveau gemeten worden.

Voor de constructie en de psychometrische analyse van de Eindtoets Basisonderwijs is het in verband met itembiasonderzoek van belang om de samenstelling en analyse in de toekomst niet op de klassieke testtheorie maar op de itemresponsetheorie te baseren. Het is dan mogelijk om allochtone en autochtone leerlingen eenduidig op basis van een eendimensionele vaardigheid in niveaugroepen in te delen, waardoor er meer duidelijkheid bestaat over de vaardigheid op basis waarvan leerlingen gematcht worden. In verband met de rapportage van de toetsuitslagen is het ook van belang om een model uit de itemresponsetheorie te gebruiken, omdat het dan mogelijk is om op minder arbitraire gronden items te schalen en te wegen. Het wegen van schalen zal nodig zijn omdat een toets die gebruikt wordt voor schoolkeuze, in verband met de predictieve validiteit, over het algehele prestatieniveau van de leerling moet rapporteren.

Hoewel tot nog toe inhoudelijk gezien de resultaten van onderzoek naar itembias bescheiden zijn, is dit soort onderzoek in verband met controle op de bruikbaarheid van evaluatie-instrumenten voor subgroepen niet overbodig. In dit verband kan opgemerkt worden dat er voorbereidingen zijn getroffen om de bestanden van de Eindtoets Basisonderwijs 1993 en van de Landelijke Evaluatie Onderwijsvoorrangsbeleid (LEO-cohort, groep acht 1993) aan elkaar te koppelen, waardoor zowel de Eindtoets Basisonderwijs 1993 als de LEO-instrumenten op toets- en itembias kunnen worden onderzocht.
De volgende aandachtspunten zijn bij toekomstig itembiasonderzoek van belang.

- Itembiasonderzoek richt zich tot nu toe voornamelijk op meerkeuze-opgaven. Er zou ook onderzoek gedaan moeten worden naar bias in open vragen.
- Itembiasonderzoek richt zich tot nu toe voornamelijk op jongens versus meisjes en op allochtone versus autochtone leerlingen. Het is te overwegen ook onderzoek te doen met andere subgroepen waarvan bekend is dat de scoreverschillen substantieel zijn, zoals leerlingen met een lage versus een hoge sociaal-economische status, leerlingen uit laag- versus hoogscorende regio's, leerlingen op effectieve versus ineffectieve scholen. Ook valt te overwegen itembiasonderzoek te doen met subgroepen die met methode (leerboek) A of met methode B onderwijs kregen.
- Itembiasonderzoek wordt voornamelijk met schoolvorderingentoetsen uitgevoerd. Er zou ook onderzoek gedaan moeten worden naar bias in tests die veelvuldig in psychologisch onderzoek gebruikt worden (vgl. Hofstee, 1990).
- De analyse-eenheid van itembiasonderzoek is meestal het afzonderlijke item. Het is te overwegen om itembiasonderzoek te doen met clusters van items. Hierbij kan gedacht worden aan de items die bij een bepaalde tekst horen, de items die dezelfde (sub)doelstelling beogen te meten of items behorende tot hetzelfde itemtype (vgl. Bügel & Glas, 1991; Dorans & Holland, 1992).
- Het is denkbaar dat cultureel bepaalde toetservaring een bron van bias is. Het gaat hierbij om de ervaring in het maken van toetsen inclusief de kennis van de soort taken die in bijvoorbeeld de Eindtoets Basisonderwijs te verwachten is. Ter voorbereiding op de toetsafname is het gebruikelijk dat leerlingen kennismaken met de itemtypen die in de toets te verwachten zijn. Er zou onderzoek gedaan kunnen worden naar de vraag hoe en hoelang allochtone maar ook autochtone leerlingen het beste op de toetsafname

kunnen worden voorbereid.
- Voor itembiasonderzoek wordt veelal een compleet, gedichotomiseerd databestand gebruikt (het item is goed of fout beantwoord). Het is te overwegen eveneens itembiasonderzoek te richten op de afleiders van het item, op de overgeslagen items of op de items waar de leerling binnen de voorgeschreven tijd niet aan toe gekomen is (Dorans & Holland, 1992; Schmitt e.a., 1992).

Summary

Suitability of the Primary School-Leaving Test for students from non-Dutch backgrounds

A description of school achievement of non-Dutch and Dutch students is often based on test achievement. Up till now, however, there has hardly been any research into the issue of whether tests which are frequently used are a proper means for measuring non-Dutch students' skills in some educational- objective areas.
Staff of the Language and Minorities Group of the Arts Faculty of the Catholic University Brabant (KUB) and staff of the Primary School-Leaving Test project of the Institute for Educational Measurement (Cito) decided to carry out a research project together to study the suitability of the Primary School-Leaving Test for non-Dutch students.
The Primary School-Leaving Test informs about school achievement of individual students, related to the choice of a school for secondary education. The test consists of 180 four-choice items, proportionally distributed over the subtests of Language, Maths and Information Processing.
The study focuses on three parts. First of all the study aims at a description of the school achievement of non-Dutch and Dutch students, i.e. the test scores of (parts of) the Primary School-Leaving Test and the data about admission to and transfer in secondary education.
The second aim is test bias. In this research context it is the study of the predictive validity of the Primary School-Leaving Test for non-Dutch and Dutch students, compared to the predictive validity of the advice given by the primary school itself.
Item bias is the focus of the third part of the study. There are two complementary phases in the item bias study: the detection and the explanation phase. In the first phase (the detection phase) statistic procedures have been used to trace biased items. There is item bias in an item when students from different subgroups with similar skills stand an unequal chance of giving the proper answer to a given item. In the second phase of the item bias study (explanation phase) an attempt has been made to study the possible cause of item bias in an item. Three groups of persons were involved in discovering the possible causes of item bias (to which purpose the items were analysed on content): the project staff (of KUB and Cito), experts not involved in the research project and students of the final grade (grade 8) of primary education. The study of test- and item bias may reveal what adjustments might enhance the suitability of the Primary School-Leaving Test for non-Dutch students.

For the study of test- and item bias five instruments have been developed. The construction of the Primary School-Leaving Test and the two questionnaires for collecting admission and transfer data in secondary education are part of the cyclic activities of the Cito Primary School-Leaving Test project. For the collection of background data at student and school level two questionnaires were especially made. The research populations consist of the students of grade eight who participated in the Primary School-Leaving Test in 1987 and 1989 respectively.

Data-analysis of the *first* part of the study, school achievement of non-Dutch and Dutch students, shows that students of Moroccan and Turkish origin had the lowest mean total score, in 1987 ánd 1989. They were followed by the students of Surinam, Antillean and Moluccan origin. Compared to all other ethnic groups (including Dutch students) the Chinese students had the highest mean maths score.
When we take a look at the students of comparable achievement level, entering secondary education, more Dutch than non-Dutch students turn out to enter types of schools with a lower average achievement level. Non-Dutch students tend to prevail in types of schools with a higher achievement level, compared to Dutch students of similar achievement levels. As the data of this present study show, the relative lead which non-Dutch students have on their Dutch counterparts at the start of their secondary education career, is nullified in part at the end of the first school year. The transfer data of students of comparable achievement level show that non-Dutch students outnumber their Dutch classmates in de-flow (they fail a year or they pass on to a type of school of a lower average achievement level). It should be noted that the differences in de-flow occur both in 1987 and in 1989, but are significant hardly anywhere.

In this thesis the study of test bias (the *second* part) is to be interpreted as checking the predictive validity of the Primary School-Leaving Test for non-Dutch and Dutch students, as compared to the advice given by the primary school itself. The scale of educational careers in secundary education, used as an dependent variable, has been determined by putting the educational positions of the students in secondary education on a scale. The predictive values of the primary school advice and the Primary School-Leaving Test have been checked by identifying the regression lines of primary school advice and Primary School-Leaving Test on the scale of educational careers for both non-Dutch and Dutch students. There appears to be a significant difference ($p<.001$), on the scale of educational careers, between the regression lines of non-Dutch and Dutch students who took part in the Primary School-Leaving Test. The difference between the regression lines, on the scale of educational careers, of non-Dutch and Dutch students based on the primary school advice is slightly significant in 1987 ($p<.05$) and non-significant in 1989.
To get a picture of the causal effect of the independent variables on the dependent variable of educational careers path-analyses have been made. These show that the school career models of both 1987 and 1989 better explain variance in educational careers in secundary education in Dutch than in non-Dutch students. In both years the primary school advice effect on educational careers outweighs the Cito-score effect. Educational careers predictions on the basis of the Cito score prove to be more accurate in non-Dutch than in Dutch students. The study also revealed more effect of the language score on educational careers in Dutch than in non-Dutch students. In non-Dutch students there is not much difference in effect between language and maths score. The effect of the information processing score is smaller in non-Dutch students.

The *third* part of the study involves the subject of item bias. In the item bias study two complementary phases have been discerned. In the first phase (the detection phase) statistic procedures are used to trace biased items. The second

phase (the explanation phase) addresses the possible causes of item bias in a given item.
There is item bias when students from different subgroups, but of a similar skill level, stand an unequal chance of giving the proper answer to the item concerned. To trace these biased items (detection phase) in the Primary School-Leaving Tests of 1987 and 1989 for Turkish and Moroccan students two item bias detection procedures have been used. The OPLM computer program (OPLM = One Parameter Logistic Model) was used as the procedure based on the Item Response Theory (IRT); the Mantel-Haenszel program was used as the procedure based on the classical test theory.
The results of the item bias analyses show that it is difficult to indicate the extent of biased items in the Primary School Leaving Tests of 1987 and 1989 for Turkish and Moroccan students. One of the Mantel-Haenszel analyses shows that the number of biased items, per subtest, varies from about fifty to twenty per cent. When uni-dimensional scale items are used in the Mantel-Haenszel procedure, 45 of the 360 analysed items turn out to be biased. Analysis of the very same uni-dimensional scales in the IRT-procedure (OPLM) reveals 20 biased items. The IRT and the Mantel-Haenszel procedure share a total number of 13 biased items. Items may be biased to the advantage or disadvantage of non-Dutch students. Further analysis reveals that both procedures (including the uni-dimensional scale items) show an eighty-seven per cent agreement in detecting (un)biased items. Comparison of the results of the various Mantel-Haenszel analyses shows a large difference in whether the items studied are of a uni- or multi-dimensional scale. The items which are biased according to the Mantel-Haenszel procedure and the items, biased from the IRT point of view, have been analysed on content and results of the content-related analyses of the two sorts of biased items have been assessed on agreement.

Tracking down possible causes of item bias (explanation phase) not only involved staff of the research project (of KUB and Cito) but also experts not involved in the project and students of the final grade (grade eight) of primary education. The items which are biased to the advantage or disadvantage of Turkish and/or Moroccan students, have been analysed first, on the basis of the skill to be measured, by the project staff, who looked for item elements which might cause this item bias. This content-oriented analysis posed two problems. On the one hand it proves to be difficult to identify, with any amount of certainty, the item element which is the source of the bias and on the other hand very similar items turn out to be item-biased one time and non-item biased another. We also found a considerable number of corresponding biased items. The results of the content analysis of items biased according to the Mantel-Haenszel and the IRT procedure show correspondence in item bias source on a number of points.
This correspondence involves the following item bias sources:
- Comprehensive reading: Items asking for an overall, a general understanding of the text may have an item bias which is beneficial to Turkish and Moroccan students. Items which ask for the meaning of a word or a sentence by demanding a literal or paraphrased repetition of information explicitly given in the text, may be item biased, to the disadvantage of these students.
- Knowledge of words and word combinations: Items asking for the meaning of difficult words, the meaning of which can hardly or with difficulty be deduced

from the context, run the risk of being item-biased, to the disadvantage of these students.
- Proper use of the language: Items involving knowledge of the form of fixed combinations of words and conventions of sentence structure stand a chance of being item-biased, to the disadvantage of Turkish and Moroccan students.
- Spelling: Items demanding identification of spelling errors in verbs and words with a fixed image, stand a chance of being item-biased, to the advantage of these students.

There is far less certainty about item bias sources in part of the clusters, as the items are only biased according to the Mantel-Haenszel procedure.

To get more indications about item bias sources experts not involved in the project have been asked to assess biased items on bias sources which are to the advantage or the disadvantage of non-Dutch students. The experts seem to have difficulties in putting their finger on items which are beneficial or non-beneficial to Turkish and Moroccan students. The correlation between the number of experts identifying an item as more difficult for non-Dutch students and item bias degree has been checked as well. Correlation between the number of experts claiming the item to be more difficult and item bias degree is not high: $r = .30$ ($p<.01$). Correspondence levels between assessments, by the experts, of item elements which are difficult for non-Dutch students turn out to be high and are generally in line with the findings of the project staff. The experts also emphasize that the culturally determined prior knowledge of non-Dutch students, which might play a part in the items involving context material, makes some items difficult for them. Non-Dutch students might be less familiar with the subject being raised in a text, for instance.

A small-scale thinking-aloud experiment was used to establish how often the wrong answer is given in biased items as the result of an item element serving as the preliminary item bias source. We also checked how often the proper answer is given in items which have been subjected to manipulation. These so called manipulated items are items in which the item element serving as the preliminary bias source (for example: 'including VAT') has been replaced by an item element which is expected not to create any bias (for example: 'with VAT'). The non-Dutch and Dutch students ticked the answer they thought right in the original and manipulated items they were presented with. Then they had to indicate as extensively and accurately as possible how they had found their answers. The students' argumentation has been examined for indications about item bias sources.
The thinking-aloud experiment indicates that bias largely seems to have its origin in the fields of word usage and implicit relations between sentences and parts of texts. Graphic obscurities may be confusing as well. Lower levels of linguistic skills in non-Dutch students may make it more difficult to handle complex items, in which students must take a number of intermediate steps to solve the item. This is particularly true when a complex item contains context material with which non-Dutch students are less familiar but which is essential in solving the item.

Literatuur

Ackerman, T.A. & J.A. Evans (1992). *An investigation of the relationship between reliability, power, and the type 1 error rate of the Mantel-Haenszel and simultaneous item bias detection procedures.* Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco: april, 1992.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education *(1985). Standards for educational and psychological tests and manuals*. Washington: American Psychological Associaton.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (ed.), *Educational measurement.* Washington: American council on education.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (ed.), *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkins University Press.

Applebee, A.N., J.A. Langer & I.V.S. Mullis (1986). *The writing report card.* Princeton: Educational Testing Service.

Baratz-Snowden, J.C. & R. Duran (1987). *The educational progress of language minority students: findings from the 1983-84 NAEP reading survey.* Princeton: Educational Testing Service.

Bergen, J.B.A.M. van (1989). *Verantwoording constructie toetsen voor de evaluatie van het onderwijsvoorrangsbeleid.* Arnhem: Instituut voor Toetsontwikkeling.

Berk, R.A. (ed.) (1982). *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Bialystok, E. & J. Cummins (1991). Language, cognition, and education of bilingual children. In E. Bialystok (ed.), *Language processing in bilingual children*. Cambridge: Cambridge University Press.

Bialystok, E. (1991). Metalinguistic dimensions of bilingual language proficiency. In E. Bialystok (ed.), *Language processing in bilingual children*. Cambridge: Cambridge University Press.

Blok, H & W.E. Saris (1980). Relevante variabelen bij het doorverwijzen na de lagere school; een structureel model. *Tijdschrift voor Onderwijsresearch*, *5*, 63-79.

Boland, T. (1991). *Lezen op termijn.* Nijmegen: Katholieke Universiteit Nijmegen.

Bosker, R.J. (1990). *Extra kansen dankzij de school?* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Bot, K. de, P. Broeder & L. Verhoeven (1985). Het meten van culturele oriëntatie in relatie met taalvaardigheid. *Toegepaste Taalwetenschap in Artikelen, 22*, 33-49.

Bügel, K (1991). Sekseverschillen in onderwijsprestaties in Nederland. *Pedagogische Studiën*, *68*, 350-370.

Bügel, K. & C. Glas (1991). Item-specifieke verschillen in prestaties van jongens en meisjes bij tekstbegripexamens moderne vreemde talen. *Tijdschrift voor Onderwijsresearch, 16*, 337-351.

Bügel, K. & H.F.M. Robben-Willems (1989). *Item-bias in examens moderne vreemde talen C/D-niveau*. Arnhem: Instituut voor Toetsontwikkeling.

Cacciari, C. & S. Glucksberg (1991). Understanding idiomatic expressions: the contribution of word meanings. In G.B. Simpson (ed.), *Understanding word and sentence*. Amsterdam: North-Holland Elsevier Science Publishers B.V.

Camilli, G. & J.K. Smith (1990). Comparison of the Mantel-Haenszel test with randomized and jackknife test for detecting biased items. *Journal of Educational Statistics, 15*, 53-67.
CBS (1986). *Leerlingen en studenten met een buitenlandse nationaliteit in het Nederlands onderwijs 1984/'85.* 's-Gravenhage: Staatsuitgeverij.
Cito (1986a). *Doelenboek, inhoudsverantwoording van de Eindtoets Basisonderwijs vanaf 1987.* Arnhem: Instituut voor Toetsontwikkeling.
Cito (1986b). *Handleiding Eindtoets Basisonderwijs 1987.* Arnhem: Instituut voor Toetsontwikkeling.
Cito (1986c). *Eindtoetsbulletin november 1986.* Arnhem: Instituut voor Toetsontwikkeling.
Cito (1988a). *Eindtoetsbulletin november 1988.* Arnhem: Instituut voor Toetsontwikkeling.
Cito (1988b). *Interpretatie van het leerlingrapport.* Arnhem: Instituut voor Toetsontwikkeling.
Cito (1990). *Interpretatie van het leerlingrapport.* Arnhem: Instituut voor Toetsontwikkeling.
Clauser, B.E., K. Mazor & R.K. Hambleton (1991). Influence of the criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. *Applied Psychological Measurement, 15*, 353-359.
Coenen, M. & T. Vallen (1991). Itembias in de eindtoets basisonderwijs. *Pedagogische Studiën, 68*, 15-26.
Cremers, P.G.J. (1980). Konstruktie van een schaal voor bereikt niveau van voortgezet onderwijs (B.N.V.O.-schaal). *Tijdschrift voor Onderwijsresearch, 5*, 80-91.
Cronbach, L.J. (1972). Judging how well a test measures. In L.J. Cronbach & P.J.D. Drenth, *Mental tests and cultural adaptation*. Den Haag: Mouton.
Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of educational Research, 49*, 222-251.
Cummins, J. (1984a). Wanted: a theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (ed.), *Language proficiency and academic achievement.* Clevedon: Multilingual Matters.
Cummins, J. (1984b). Language proficiency and academic achievement revisited: a response. In C. Rivera (ed.), *Language proficiency and academic achievement*. Clevedon: Multilingual Matters.
Cummins, J. (1991a). Language development and academic learning. In L. Malavé & G. Duquette (eds.), *Language, culture and cognition. A collection of studies in first and second language acquisition*. Clevedon: Multilingual Matters.
Cummins, J. (1991b). Interdependence of first- and second-language proficiency in bilingual children. In E. Bialystok (ed.), *Language processing in bilingual children.* Cambridge: Cambridge University Press.

Diaz, R.M. & C. Klingler (1991). Towards an explanatory model of the interaction between bilingualism and cognitive development. In E. Bialystok (ed.), *Language processing in bilingual children*. Cambridge: Cambridge University Press.
Dorans, N.J. & P.W. Holland (1992). *DIF detection and description: Mantel-Haenszel and standardization.* Princeton: Educational Testing Service.
Dossey, J.A., I.V.S. Mullis, M.M. Lindquist & D.L. Chambers (1988). *The Mathematics Report Card.* Princeton: Educational Testing Service.
Drenth, P.J.D. (1972). Implications of testing for individual and society. In L.J. Cronbach & P.J.D. Drenth, *Mental tests and cultural adaptation*. Den Haag: Mouton.

Drenth, P.J.D. (1973). *De psychologische test.* Deventer: Van Loghem Slaterus.
Driessen, G. (1990). *De onderwijspositie van allochtone leerlingen.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Driessen, G. (1991a). Discrepanties tussen toetsresultaten en doorstroomniveau. Positieve discriminatie bij de overgang basisonderwijs - voortgezet onderwijs? *Pedagogische Studiën, 68*, 27-35.
Driessen, G. (1991b). Marokkaanse kinderen op Nederlandse scholen, een exploratie van hun achtergronden en onderwijsprestaties. *Tijdschrift voor Onderwijswetenschappen, 21*, 292-306.

Ekstrom, R.B., M.E. Lockheed & T.F. Donlon (1979). Sex differences and sex bias in test content. *Educational Horizons, 47*-52.
Engelen, R.J.H. & J.H. Uiterwijk (1990). *Verantwoording Eindtoets Basisonderwijs 1987.* Arnhem: Instituut voor Toetsontwikkeling.
Engelhard, G., L. Hansche & K.E. Rutledge (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*, 347-360.
Ersoy, D. (1991). Cito-toetsen en schoolkeuze allochtonen. *Het Schoolblad, 14*, 42.
Esch, W. van (1983). *Toetsprestaties en doorstroomadviezen van allochtone leerlingen in de zesde klas van lagere scholen.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Essen, E. van (1983). *De invloed van de eindtoetsscore op het schoolkeuze-advies van het schoolhoofd en de toelating van leerlingen tot het voortgezet onderwijs.* Arnhem: Instituut voor Toetsontwikkeling.
Extra, G. & T. Vallen (1985). Languages and ethnic minorities in the Netherlands: Current issues and research areas. In G. Extra & T. Vallen (eds.), *Ethnic minorities and Dutch as a second language.* Dordrecht: Foris Publications.
Extra, G. & L. Verhoeven (1985). Bias in intelligentie-onderzoek bij allochtone kinderen. *Pedagogische Studiën, 62*, 392-395.
Extra, G. & L. Verhoeven (1990). *Ethnic minority research in the Netherlands, crosslinguistic perspectives on Turkish and Moroccan communities abroad.* Paper presented at the International Workshop on Ethnic Community Languages in Europe. Gilze-Rijen: december, 1990.

Genesee, F. (1984). On Cummins' theoretical framework. In C. Rivera (ed.), *Language proficiency and academic achievement.* Clevedon: Multilingual Matters.
Gernsbacher, M.A. & M. Faust (1991). The role of suppression in sentence comprehension. In G.B. Simpson (ed.), *Understanding word and sentence.* Amsterdam: North-Holland Elsevier Science Publishers B.V.
Glas, C.A.W. (1991). *Testing Rasch models for polytomous items.* Arnhem: Instituut voor Toetsontwikkeling.
Glas, C.A.W. & M.J. Ouborg (1993). Vraagonzuiverheid. In T.J.H.M. Eggen & P.F. Sanders (eds.), *Psychometrie in de praktijk.* Arnhem: Instituut voor toetsontwikkeling.
Glas, C.A.W. & N.D. Verhelst (1993). Een overzicht van itemresponsmodellen. In T.J.H.M. Eggen & P.F. Sanders (eds.), *Psychometrie in de praktijk.* Arnhem: Instituut voor toetsontwikkeling.
Groot, A.D. de & R.F. van Naerssen (eds.) (1969). *Studietoetsen construeren, afnemen en analyseren.* Den Haag: Mouton.

Hacquebord, H. (1989). *Tekstbegrip van Turkse en Nederlandse leerlingen in het voortgezet onderwijs.* Dordrecht: Foris Publications.
Hakuta, K. (1986). *Mirror of language. The debate on bilingualism.* New York: Basic Books, Inc., Publishers.
Hambleton, R.K. & H.J. Rogers (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.
Hambleton, R.K. & R.W. Jones (1992). *Comparison of empirical and judgmental methods for detecting differential item functioning.* Princeton: Educational Testing Service.
Hills, J.R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice*, 5-11.
Hoeven-van Doornum, A.A. van der (1990). *Effecten van leerlingbeelden en streefniveaus op schoolloopbanen.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Hof, L. van 't & J. Dronkers (1992). *Onderwijsachterstanden van allochtonen: klasse, gezin of cultuur.* Amsterdam: Stichting Centrum voor Onderwijsonderzoek.
Hofstee, W.K.B. (1990). Toepasbaarheid van psychologische tests bij allochtonen. *De Psycholoog, 25*, 291-294.
Holland, P.W. & H. Wainer (eds.) (1993). *Differential item functioning.* Hillsdale: Lawrence Erlbaum Associates.
Holland, P.W. & D.T. Thayer (1986). *Differential item performance and the Mantel-Haenszel procedure.* Paper presented at the American Educational Research Association Annual Meeting. San Francisco: april, 1986.

Intraprasert, D. (1986). *An investigation of the reliability of five methods for detecting test item bias: an empirical study.* Denton: North Texas State University.
Ironson, G.H. (1982). Use of Chi-square and latent trait approaches for detecting item bias. In R.A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Jensen, A.R. (1980). *Bias in mental testing.* Londen: Methuen.
Johnson, J. (1991). Constructive processes in bilingualism and cognitive growth effects. In E. Bialystok (ed.), *Language processing in bilingual children.* Cambridge: Cambridge University Press.
Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research quarterly, 19*, 219-240.
Jong, M. de, H. Uiterwijk, A. Kerkhoff & T. Vallen (1987). *Vooronderzoek naar item- en testbias in de Eindtoets Basisonderwijs.* Tilburg: Faculteit der Letteren.
Jong, M. de & T. Vallen (1989). Linguïstische en culturele bronnen van itembias in de Eindtoets Basisonderwijs voor leerlingen uit etnische minderheidsgroepen. *Pedagogische Studiën, 66*, 390-402.
Jong, M.J. de (1987). *Herkomst, kennis en kansen.* Lisse: Swets & Zeitlinger.
Jungbluth, P. (1985). *Verborgen differentiatie.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Jungbluth, P. & A. van Langen (1990). Onderwijsondersteunend thuisklimaat bij migranten. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving*. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Jungbluth, P., A. van Langen & H. Vierke (1990). De trefzekerheid van het advies voortgezet onderwijs bij migranten. *Tijdschrift voor Onderwijswetenschappen, 21*, 90-101.

Kerkhoff, A. (1988). *Taalvaardigheid en schoolsucces.* Lisse: Swets & Zeitlinger.
Kerkhoff, A. & T. Vallen (1985). Cultural biases in second language testing of children. In G. Extra & T. Vallen (eds.), *Ethnic minorities and Dutch as a second language.* Dordrecht: Foris Publications.
Klaassen, C.A.C. & P.L.M. Jungbluth (1990). *Onderwijs research dagen 1990, onderwijs en samenleving.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Klerk, L.F.W. de (1983). *Onderwijspsychologie.* Deventer: Van Loghum Slaterus.
Kok, F. (1988). *Vraagpartijdigheid.* Amsterdam: Universiteit van Amsterdam.
Koopman, P., P. van den Eeden & U. de Jong (1986). Categoriale MAVO-scholen en schoolloopbanen in Amsterdam. In W.J. Nijhof & E. Warries (eds.), *De opbrengst van onderwijs en onderzoek.* Lisse: Swets & Zeitlinger.
Kreft, I & J. de Leeuw (1986). *Maken scholen verschil?* Paper voor de gezamenlijke bijeenkomst van de werkgroepen Longitudinaal School- en Beroepsloopbaanonderzoek en Multilevel Onderzoek. Leiden: Rijksuniversiteit Leiden.
Kuhlemeier, H. & H. van den Bergh (1991). De correlationele structuur van taalvaardigheid: een exploratie. *Tijdschrift voor Onderwijsresearch, 16*, 141-159.

Lalleman, J. (1986). *De invloed van sociale en sociaal-psychologische factoren op T2-vaardigheid.* Amsterdam: Universiteit van Amsterdam.
Lange, R. de & J. Rupp (1990). Tegenstrijdigheden in de schoolloopbanen van Turkse en Marokkaanse leerlingen. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Langen, A. van & P. Jungbluth (1990). *Onderwijskansen van migranten.* Forum 6. Amsterdam/Lisse: Swets & Zeitlinger.
Langen, A. van & P. Jungbluth (1992). Leerkracht-overwegingen bij de vervolgadvisering van migranten. *Tijdschrift voor Onderwijswetenschappen, 22*, 69-79
Linn, R.L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick, *Principals of modern psychological measurement.* New Jersey: Lawrence Erlbauw Ass.

Malpass, R.S. & Y.H. Poortinga (1986). Strategies for design and analyses. In W.J. Lonner & J.W. Berry (eds.), *Field methods in cross-cultural research.* Beverly Hills: Sage Publications.
Maureau, J.H. (1979). *Goed en begrijpelijk schrijven. Een analyse van 40 jaar schrijfadviezen.* Muiderberg: Coutinho.
Meijnen, G.W. (1979). *Maatschappelijke achtergronden van intellektuele ontwikkeling.* Groningen: Wolters-Noordhoff.
Meijnen, G.W. (1984). *Van zes tot twaalf. Een longitudinaal onderzoek naar milieu- en schooleffecten van loopbanen in het lager onderwijs.* Harlingen: Flevodruk Harlingen.
Meijnen, G.W. & F.S.J. Riemersma (1992). Schoolcarrières: een klassenkwestie? In Commissie Allochtone Leerlingen in het Onderwijs, *Ceders in de tuin, deel twee.* Zoetermeer: Ministerie van Onderwijs en Wetenschappen.
Mellenbergh, G.J. (1989). Itembias and itemresponse theory. In R.K. Hambleton (ed.), Applications of itemresponse theory (special issue). *International Journal of Educational Research, 13*, 127-143.
Messick, S. (1986). *The once and future issues of validity: assessing the meaning and consequences of measurement.* Princeton: Educational Testing Service.
Messick, S. (1987). *Validity.* Princeton: Educational Testing Service.
Mulder, L. (1993). *De tweede fase van de OVB-cohortonderzoeken in het basisonderwijs.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Mulder, L. & B. Pijl (1992). *De onderwijspositie van leerlingen uit de OVB-doelgroepen na twee jaar voortgezet onderwijs.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Mulder, L. & P. Tesser (1990). Cultureel kapitaal en schoolprestaties in het basisonderwijs. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving*. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Mulder, L. & P. Tesser (1991). *De schoolkeuze van allochtone leerlingen*. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Nienhuis, L.J.A. (1991). Het begrip van teksten met 10% tot 25% onbekende woorden. *Toegepaste taalwetenschap in artikelen, 41*, 57-66.

O'Malley, J.M. & A.U. Chamot (1990). *Learning strategies in second language acquisition.* Cambridge: Cambridge University Press.

Rayner, K. & R.K. Morris (1991). Comprehension processes in reading ambiguous sentences: reflections from eye movements. In G.B. Simpson (ed.), *Understanding word and sentence*. Amsterdam: North-Holland Elsevier Science Publishers B.V.
Raju, N.S., R.K. Bode & V.S. Larsen (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. *Applied measurement in education, 2*, 1-13.
Reezigt, G. & M. Weide (1990). Determinanten van achterstanden bij allochtone leerlingen in het basisonderwijs. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving*. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Reynolds, C.R. (1982). Methods for detecting construct and predictive bias. In R.A. Berk (ed.), *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkins University Press.
Reynolds, R.E., M.A. Taylor, M.S. Stefensen, L.L. Shirey & R.C. Anderson (1982). Cultural schemata and reading comprehension. *Reading Research Quarterly, 17*, 353-366.
Roelandt, T., E. Martens & J. Veenman (1990). Achterstand van allochtonen in het onderwijs: sociaal milieu en migratie-achtergronden. *Mens en Maatschappij, 65*, 103-125.

Scheuneman, J.D. (1982). A posteriori analyses of biased items. In R.A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.
Scheuneman, J.D. (1985). *Exploration of causes of bias in test items.* Princeton: Educational Testing Service.
Scheuneman, J.D. (1988). Item bias and individual differences. In S. Irvine (ed.), *Human assessment in computer context.* Den Haag: Nijhoff.
Scheuneman, J.D. & K.S. Steinhaus (1987). *A theoretical framework for the study of item difficulty and discrimination.* Princeton: Educational Testing Service.
Scheuneman, J.D. & K. Gerritz (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*, 109-131.
Schmitt, A.P., P.W. Holland & N.J. Dorans (1992). *Evaluating hypotheses about differential item functioning.* Princeton: Educational Testing Service.
Schwanenflugel, P.J. (1991). Contextual constraint and lexical processing. In G.B. Simpson (ed.), *Understanding word and sentence.* Amsterdam: North-Holland Elsevier Science Publishers B.V.

Shealy, R.T. & W.F. Stout (1993). An item response theory model for test bias and differential test functioning. In P.W. Holland & H.W. Wainer (eds.), *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.
Shepard, L.A. (1982). Definitions of bias. In R.A. Berk (ed.), *Handbook of methods for detecting test bias*. Baltimore: The Johns Hopkins University Press.
Sijtstra, J. (1991). *Doel en inhoud van taalonderwijs. De ontwikkeling van een model voor domeinbeschrijvingen van taalonderwijs.* Arnhem: Instituut voor toetsontwikkeling.
Skaggs, G. & R.W. Lissitz (1988). *Consistency of selected item bias indices.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans: april 1988.
Sman, J.H.A. van der & J.H. Uiterwijk (1985). *Verantwoording Eindtoets Basisonderwijs 1983.* Arnhem: Instituut voor Toetsontwikkeling.
Spaeth, J.L. (1975). Path analysis. In D.J. Amick & H.J. Walberg (eds.), *Introductory multivariate analysis*. Berkeley: Mc. Cutchan Publishing Corp.
Swinney, D.A. (1991). The resolution of indeterminacy during language comprehension: perspectives on modularity in lexical, structural and pragmatic processing. In G.B. Simpson (ed.), *Understanding word and sentence*. Amsterdam: North-Holland Elsevier Science Publishers B.V.

Tabossi, P. (1991). Understanding words in context. In G.B. Simpson (ed.), *Understanding word and sentence*. Amsterdam: North-Holland Elsevier Science Publishers B.V.
Tatsuoka, K.K., R.L. Linn, M.M. Tatsuoka & K. Yamamoto (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25*, 301-319.
Taylor, I. & M.M. Taylor (1990). *Psycholinguistics. Learning and using language.* New Jersey: Prentice Hall, Inc.
Tesser, P. (1986). *Sociale herkomst en schoolloopbanen in het voortgezet onderwijs.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Tesser, P. & L. Mulder (1990). Cultureel kapitaal en schoolprestaties in het basisonderwijs. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving*. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Tesser, P. & H. Vierke (1990). *De schoolprestaties van allochtone leerlingen in het basisonderwijs.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Tesser, P., L. Mulder & G. van der Werf (1991). *De eerste fase van de longitudinale OVB-onderzoeken, het leerlingenonderzoek.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Uiterwijk, J.H. (1990a). *Item- en testbias in de Eindtoets Basisonderwijs 1987.* Arnhem: Instituut voor Toetsontwikkeling.
Uiterwijk, J.H. (1990b). Verschillen tussen autochtonen en allochtonen bij de overgang van basisonderwijs naar voortgezet onderwijs. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving*. Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Uiterwijk, J.H. & R.J.H. Engelen (1991). *Verantwoording Eindtoets Basisonderwijs 1988.* Arnhem: Instituut voor Toetsontwikkeling.
Uiterwijk, J.H. & R.J.H. Engelen (1992). *Verantwoording Eindtoets Basisonderwijs 1989.* Arnhem: Instituut voor Toetsontwikkeling.

Uiterwijk, J.H. & T. Vallen (1991). De bruikbaarheid van de Cito-Eindtoets Basisonderwijs voor leerlingen uit etnische minderheidsgroepen; een eerste analyse. In R. van Hout & E. Huls (eds.), *Artikelen van de Eerste Sociolinguïstische Conferentie*. Delft: Eburon.

Vallen, T. & A. Kerkhoff (1985). Beheersing van het Nederlands en doorstroming lager onderwijs/voortgezet onderwijs bij kinderen uit etnische minderheidsgroepen. In W.K.B. Koning (ed.), *Taalbeheersing in Theorie en Praktijk*. Dordrecht: Foris Publications.
Velden, R.K.W. van der (1991). *Sociale herkomst en schoolsucces.* Monografieën onderwijsonderzoek nr. 10. Groningen: Instituut voor Onderwijsonderzoek.
Verhelst, N.D. (1988). *De Mantel-Haenszel-toetsen.* Arnhem: Instituut voor Toetsontwikkeling.
Verhelst, N.D. (1992). *Het eenparameter logistisch model (OPLM), een theoretische inleiding en een handleiding bij het computerprogramma.* Arnhem: Instituut voor toetsontwikkeling.
Verhoeven, L. & A. Vermeer (1992). Woordenschat van leerlingen in het basis- en MLK-onderwijs. *Pedagogische Studiën, 69*, 218-234.
Verhoeven, L. (1987). *Ethnic minority children acquiring literacy.* Dordrecht: ICG.
Vermeer, A. (1986). *Tempo en struktuur van tweede-taalverwerving bij Turkse en Marokkaanse kinderen*. Tilburg: KUB
Visser, J.J.C.M. & M.J.M. Voeten (1987). *Het advies voor voortgezet onderwijs.* Nijmegen: Katholieke Universiteit Nijmegen.
Vijver, F. van de (1991). *Inductive thinking across cultures: an empirical investigation.* Helmond: Wibro
Vijver, F. van de & Y.H. Poortinga (1991). Testing across cultures. In R.K. Hambleton & J.N. Zaal (eds.), *Advances in educational and psychological testing*. Dordrecht: Kluwer Academic Publishers.
Vijver, F. van de, G. Willemse & B. van de Rijt (1993). Het testen van cognitieve vaardigheden van allochtone leerlingen. *De Psycholoog, 28*, 152-159.

Waal, M. van de (1992). *Expert-oordelen over potentiële bronnen van itembias in de Eindtoets Basisonderwijs.* Tilburg: Katholieke Universiteit Brabant.
Wald, B. (1984). A sociolinguistic perspective on Cummins' current framework for relating language proficiency to academic achievement. In C. Rivera (ed.), *Language proficiency and academic achievement.* Clevedon: Multilingual Matters.
Weide, M. & G. van der Werf (1990). Allochtone leerlingen en hun ouders: de rol van onderwijsondersteunend gedrag. In C.A.C. Klaassen & P.L.M. Jungbluth (eds.), *Onderwijs research dagen 1990, onderwijs en samenleving.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Westerlaak, J.M., J.A. Kropman & J.W.M. Collaris (1975). *Beroepenklapper.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.
Whitney, P. & D.A. Waring (1991). The role of knowledge in comprehension: a cognitive control perspective. In G.B. Simpson (ed.), *Understanding word and sentence.* Amsterdam: North-Holland Elsevier Science Publishers B.V.
Wijnstra, J.M. (1984a). *Verantwoording Eindtoets Basisonderwijs 1981.* Arnhem: Instituut voor Toetsontwikkeling.
Wijnstra, J.M. (1984b). *Verantwoording Eindtoets Basisonderwijs 1982.* Arnhem: Instituut voor Toetsontwikkeling.
Wijnstra, J.M. (ed.) (1988). *Balans van het rekenonderwijs in de basisschool.* Arnhem: Instituut voor Toetsontwikkeling.

Wit, W. de, C. Suhre & L. Mulder (1993). *De onderwijspositie van de OVB-doelgroepleerlingen na drie jaar voortgezet onderwijs.* Nijmegen: Instituut voor Toegepaste Sociale Wetenschappen.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H.W. Wainer (eds.), *Differential item functioning*. Hillsdale: Lawrence Erlbaum Associates.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185-197.

Bijlagen

Bijlage 1: Vragenlijst op leerlingniveau (Vragenlijst B)

Vragenlijst B

- Deze vragenlijst bevat zes vragen op leerlingniveau. Wij verzoeken u deze vragen voor alle autochtone èn alle allochtone leerlingen die aan de Cito-Eindtoets-Basisonderwijs 1989 deelnemen, te beantwoorden. De antwoorden vult u in op de bijgevoegde antwoordbladen aan de hand van de bij iedere vraag gegeven toelichting.

- Alvorens u met de beantwoording van de leerlingvragen begint, vult u telkens eerst de achternaam en de voorletters van de betreffende leerling in. Wilt u erop letten dat de naam van de leerling en het leerlingnummer corresponderen met de naam van de leerling en het leerlingnummer op de antwoordbladen voor de Cito-Eindtoets 1989. U kunt dus het beste telkens de gegevens van de antwoordbladen voor de Eindtoets 1989 overnemen.

- De antwoordbladen zullen op het Cito machinaal worden verwerkt. In verband hiermee is het noodzakelijk dat de van toepassing zijnde merkplaats met potlood wordt ingevuld.

- Per vraag mag slechts één antwoordmogelijkheid aangestreept worden.

Vragen op leerlingniveau (voor alle Eindtoets-deelnemers beantwoorden)

1 *Wat is het herkomstland van beide ouders/verzorgers van de leerling?*

Toelichting:

Een leerling wordt ingedeeld bij één van de hieronder te noemen antwoordmogelijkheden, wanneer het herkomstland van beide ouders/verzorgers overeenstemt met één van deze antwoordmogelijkheden.

Bij één-ouder-gezinnen geldt het herkomstland van de ouder/verzorger bij wie het kind woont als criterium om vraag 1 te beantwoorden.

Een kind valt onder 'Overige' wanneer:
- (bij twee-ouder-gezinnen) de ouders/verzorgers twee verschillende herkomstlanden hebben;
- (bij twee-ouder-gezinnen) beide ouders/verzorgers een ander herkomstland hebben dan de hieronder genoemde (bijv. Australië);
- (bij één-ouder-gezinnen) de ouder/verzorger bij wie het kind woont een ander herkomstland heeft dan de hieronder genoemde (bijv. Nieuw-Zeeland).

N.B.: U zult merken dat wij een indeling hanteren die afwijkt van de indeling in de Formatieregeling WBO. De door ons gehanteerde indeling wordt alleen gebruikt voor onze onderzoeksdoeleinden, en staat geheel los van de indeling van het Ministerie van Onderwijs en Wetenschappen.

U kunt kiezen uit de volgende antwoordmogelijkheden:

- Nederland
- Turkije
- Marokko
- Zuid-Europa (Italië, Spanje, Portugal incl. Kaapverdië, Griekenland, Joegoslavië)
- Oost-Europa (D.D.R., Polen, Tsjecho-Slowakije, Hongarije, U.S.S.R., Roemenië, Bulgarije, Albanië)
- Noord- en West-Europa (alle overige Europese landen, exclusief Nederland)
- China* (Volksrepubliek China, Taiwan, Hongkong, Singapore)
- Molukken*
- Nederlandse Antillen
- Suriname: Creolen
- Suriname: Hindoestanen
- Overige (zie hierboven)

* Omdat met name bij Molukkers en Chinezen ook tweede-generatie-kinderen voorkomen, wordt een leerling óók tot deze categorie gerekend, wanneer de grootouders het betreffende herkomstland hebben.

Het antwoord op deze vraag vult u in achter 1 op het antwoordblad.

2 *In welk leerjaar startte de leerling met het volgen van onderwijs aan een Nederlandse basisschool?*

Toelichting:

Deze vraag heeft betrekking op de totale schoolloopbaan van de leerling in Nederland (dus niet alleen op die aan uw eigen school).
Omdat de meeste leerlingen veelal vóór 1 augustus 1985 voor het eerst in het Nederlandse onderwijs zijn begonnen, hanteren wij op het antwoordblad de oude benamingen k.o. (kleuteronderwijs) en l.o. (lager onderwijs).

Het antwoord op deze vraag vult u in achter 2 op het antwoordblad.

3 *Hoe vaak heeft de leerling tot nu toe gedoubleerd?*

Toelichting:

Het antwoord op deze vraag vult u in achter 3 op het antwoordblad.

4 *In welke mate bent u het eens met onderstaande uitspraak over de leerling? "Deze leerling heeft een groot abstractievermogen".*

Toelichting:

Deze vraag heeft betrekking op één aspect van de leerling: het abstractievermogen. Daaronder verstaan wij het vermogen om hoofd- en bijzaken te onderscheiden en de zaken snel te doorzien, waardoor goed en snel inzicht wordt verkregen.

Het antwoord op deze vraag vult u in achter 4 op het antwoordblad (de vijf antwoordmogelijkheden lopen van 'oneens' tot 'eens').

5 *In welke mate bent u het eens met onderstaande uitspraak over de leerling? "Het sociaal-culturele klimaat in het gezin biedt de leerling een goede kans op een succesvolle schoolloopbaan in het Nederlandse onderwijssysteem".*

Toelichting:

Ook deze vraag moet voor zowel alle autochtone als alle allochtone leerlingen ingevuld worden.
Het antwoord vult u in achter 5 op het antwoordblad (de vijf antwoordmogelijkheden lopen van 'oneens' tot 'eens').

6 *Voor welk type van vervolgonderwijs is de leerling naar uw oordeel het meest geschikt?*

Toelichting:

Het gaat bij deze vraag om het onderwijsniveau waarvoor de leerling naar uw eigen oordeel het meest geschikt is. De combinatie van twee typen vervolgonderwijs (IBO/LBO, LBO/MAVO, MAVO/HAVO, HAVO/VWO) dient u alleen te kiezen bij echte twijfelgevallen.

U kunt kiezen uit de volgende antwoordmogelijkheden:

- Speciaal onderwijs (speciaal basisonderwijs of speciaal voortgezet onderwijs)
- Basisonderwijs: doubleren groep acht
- IBO
- IBO/LBO
- LBO
- LBO/MAVO
- MAVO
- MAVO/HAVO
- HAVO
- HAVO/VWO
- VWO

Het antwoord op deze vraag vult u in achter 6 op het antwoordblad.

Wij danken u hartelijk voor uw medewerking.

Bijlage 2: Vragenlijst op schoolniveau (Vragenlijst A)

Vragenlijst A

Deze vragenlijst bevat twee vragen. De eerste vraag heeft betrekking op uw hele school (groep 1 t/m 8). De tweede vraag heeft betrekking op alle leerlingen in groep 8 van uw school.

Vraag op schoolniveau

1 *Wat was op 16 januari 1989 het aantal leerlingen per achterstandscategorie op uw school, alsmede het totaal aantal leerlingen op uw school?*

Toelichting:

Het antwoord op deze vraag is af te lezen uit de tabel bij vraag 1 van het telformulier van 16 januari 1989, dat uw school naar de inspectie heeft gezonden. U hoeft alleen de onderste regel over te nemen met de totale aantallen jongens en meisjes per achterstandscategorie én het totale aantal leerlingen. Deze aantallen vult u hieronder in.

Vraag 1 Aantal leerlingen naar leeftijd, achterstandscategorie en geslacht

	Achterstandscategorie / geslacht										
	1.00		1.25		1.40		1.70		1.90		Totaal
Leeftijd	J	M	J	M	J	M	J	M	J	M	
4 jarigen											
5 jarigen											
6 jarigen											
7 jarigen											
8 jarigen											
9 jarigen											
10 jarigen											
11 jarigen											
12 jarigen											
13 jarigen											
14 en ouder											
Totaal											

Vraag op groepsniveau

2 *Hoeveel leerlingen uit de totale groep 8 van uw school (parallelklassen samen nemen!) doen dit jaar niet mee aan de Cito-Eindtoets-Basisonderwijs, omdat ze de Nederlandse taal niet voldoende beheersen om de opgaven te kunnen lezen?*	vraag 2

Toelichting:

Bij deze vraag gaat het om de opgave van het totale aantal leerlingen waarop de volgende passage op blz. 9 van de Handleiding Cito-Eindtoets-Basisonderwijs 1989 van toepassing is:
"We moeten een uitzondering maken voor de kinderen die de Nederlandse taal niet voldoende beheersen om de opgaven te kunnen lezen. Dit zijn in de regel kinderen van buitenlandse werknemers die nog niet zo lang in Nederland zijn. Behalve deze groep doen alle leerlingen van groep acht mee".

Het aantal leerlingen uit de totale groep 8 van uw school (parallelklassen optellen) dat op grond van deze bepaling niet aan de Cito-Eindtoets-Basisonderwijs meedoet, vult u hierboven in in het hokje achter vraag 2.

Eindtoets Basisonderwijs

Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs
Eindtoets Basisonderwijs